Journal of Cheminformatics

RESEARCH Open Access

Improving the accuracy of prediction models for small datasets of Cytochrome P450 inhibition with deep learning

Elpri Eka Permadi[1,2], Reiko Watanabe[1*] and Kenji Mizuguchi[1*]

Abstract

The cytochrome P450 (CYP) superfamily metabolises a wide range of compounds; however, drug-induced CYP inhibition can lead to adverse interactions. Identifying potential CYP inhibitors is crucial for safe drug administration. This study investigated the application of deep learning techniques to the prediction of CYP inhibition, focusing on the challenges posed by limited datasets for CYP2B6 and CYP2C8 isoforms. To tackle these limitations, we leveraged larger datasets for related CYP isoforms, compiling comprehensive data from public databases containing IC50 values for 12,369 compounds that target seven CYP isoforms. We constructed single-task, fine-tuning, multitask, and multitask models incorporating data imputation on the missing values. Notably, the multitask models with data imputation demonstrated significant improvement in CYP inhibition prediction over the single-task models. Using the most accurate prediction models, we evaluated the inhibitory activity of approved drugs against CYP2B6 and CYP2C8. Among the 1,808 approved drugs analysed, our multitask models with data imputation identified 161 and 154 potential inhibitors of CYP2B6 and CYP2C8, respectively. This study underscores the significant potential of multitask deep learning, particularly when utilising a graph convolutional network with data imputation, to enhance the accuracy of CYP inhibition predictions under the conditions of limited data availability.

Scientific contribution

This study demonstrates that even with small datasets, accurate prediction models can be constructed by utilising related data effectively. Also, our imputation techniques on the missing values improved the prediction accuracy of CYP2B6 and CYP2C8 inhibition significantly.

Keywords Cytochrome P450, Deep learning, Small dataset, Multitask, Fine-tuning, Missing values, Imputation

Introduction

Human cytochrome P450 (CYP) represents a family of enzymes with 57 isoforms that are responsible for the biotransformation of endogenous and exogenous compounds, including drugs and toxins, via oxidation and reduction. These membrane-attached haemoprotein enzymes are primarily found in the smooth endoplasmic reticulum and are predominantly associated with hepatic cells [1, 2]. About 15 isoforms belong to CYP families 1, 2, and 3 (70–80% of all Phase I metabolisms of clinically used drugs) and are involved in the biotransformation of environmental chemicals (approximately 90%), including

*Correspondence:
Reiko Watanabe
reiko-watanabe@protein.osaka-u.ac.jp
Kenji Mizuguchi
kenji@protein.osaka-u.ac.jp
[1] Laboratory for Computational Biology, Institute for Protein Research, Osaka University, 3-2 Yamadaoka, Suita, Osaka 565-0871, Japan
[2] Research Center for Pharmaceutical Ingredients and Traditional Medicine, National Research and Innovation Agency (BRIN), Meatpro Building, Kawasan Sains dan Teknologi (KST) Soekarno, Jl. Raya Jakarta-Bogor KM 46, Cibinong, Jawa Barat 16911, Indonesia

66% of the metabolic reactions of chemical carcinogens [3]. The major isoforms that metabolise over 90% of drugs are CYP1A2, CYP2A6, CYP2B6, CYP2C8, CYP2C9, CYP2C19, CYP2D6, CYP2E1, and CYP3A4 [4, 5].

In clinical practice, patients are often prescribed multiple drugs or a combination thereof for treatment. Although this may represent a preferred clinical strategy, the administration of drug combinations may cause undesirable drug-drug interactions (DDI). As the risk of DDI increases, it can cause side effects and even increase the possibility of severe adverse effects. From a pharmacokinetic viewpoint, DDI occur when one drug alters the disposition of another co-administered drug, either increasing or decreasing its activity [6]. Therefore, inferring the possibility of DDI by understanding CYP inhibition activity is highly advantageous in the early stages of drug development to reduce the occurrence of underperforming drug candidates.

Currently, the availability of experimental data on chemical-CYP interactions is increasing, and many researchers have used computational approaches to predict or explore CYP-mediated metabolism and inhibition. However, it is difficult to accurately predict CYP450 inhibitors using structure-based techniques such as molecular docking and pharmacophore mapping because of the flexible conformation of CYP450 [7]. In contrast, machine learning is the most popular approach for quantitative structure–activity relationships (QSAR), and is widely used to predict CYP450 inhibitors [8]. Previous studies have attempted to predict CYP inhibitors using different machine learning approaches with varying accuracies [5, 9–12]. Considering the sequence and structural similarities of the binding sites in the CYP family [13], multitask models can simultaneously predict the inhibitors of different CYP isoforms to provide better predictive power [8]. Li et al. [7] constructed a multitask learning model using deep autoencoder neural networks for five major CYP enzymes (CYP1A2, CYP2D6, CYP2C9, CYP2C19, and CYP3A4) and concluded that multitask models tend to improve the performance compared to single-task models. Similar results were reported by Nguyen-Vo et al. [1], who developed iCYP-MFE models that combine multitask learning with molecular fingerprint-embedded encoding. Their models improved prediction performance slightly over the Swiss-ADME and SuperCYP models. The latest study on multitask learning was conducted by Ai et al. [8], who showed that their approach using a fingerprint-based graph neural network architecture (named DEEPCYPs) can slightly improve their CYP inhibitor prediction over iCYP and SuperCYP models. While multitask learning can enhance the performance of CYP inhibitor prediction, most of the previous research has focused on only five major isoforms, and not on other related CYPs, such as CYP2B6 and CYP2C8.

The functions of these two isoforms are relevant. CYP2B6 is involved in the metabolism of approximately 7% of clinical drugs, including the psychiatric drug mephobarbital, the antidepressant bupropion, the anaesthetic drugs profol and amiodone, the anti-cancer drug cyclophosphamide, and the anti-viral efarivens [4, 14]. In addition, both contribute to the metabolism of *n*-hexane and monoterpenes. CYP2C8 accounts for approximately 6–7% of the total hepatic CYP content and contributes to the metabolism of paclitaxel, amodiaquine, rosiglitazone, and flurbiprofen. Both also contribute to metabolising fatty acids [14, 15]. The U.S. Food and Drug Administration (FDA) on their guidelines encouraged CYP-based DDI studies for CYP2B6 and CYP2C8 in 2012 and 2006, respectively; however, the measured inhibition data were limited due to these CYPs being added later than other major CYP isoforms [16, 17]. Indeed, the amount of experimental inhibition data for CYP2B6 and CYP2C8 is severely limited in public databases, such as ChEMBL and PubChem. Building a predictive model for small datasets, such as CYP2B6 or CYP2C8, is challenging; the small amounts of data and imbalances tend to cause model overfitting or underfitting because of the small data scale and feature dimensions that are too high or low [18]. In addition, the inhibitory activities of many approved drugs against CYP2B6 and CYP2C8 remain unknown. A comprehensive prediction of the inhibitory activity of approved drugs against CYP2B6 and CYP2C8 can help identify potential inhibitors, which would be beneficial for ensuring their safety after marketing.

In this study, we constructed prediction models for CYP inhibitory activity on small dataset of CYP2B6 and CYP2C8 using an extensive dataset obtained by comprehensively collecting and integrating public data on CYP inhibition. Furthermore, we leveraged these prediction models to predict the CYP inhibitory activity of the approved drugs on CYP2C8 and CYP2B6 and identified compounds with potential CYP inhibitory activity.

Results and discussion

Dataset construction

A total of 170,355 data points of IC50 values for CYP1A2, CYP2B6, CYP2C8, CYP2C9, CYP2C19, CYP2D6, and CYP3A4 were obtained from ChEMBL [19], PubChem [20], and Rudik et al. [21]. After curation, we constructed a high/low labelled dataset consisting of 12,369 compounds with a threshold of pIC50 = 5 (IC50 = 10 μM); further details of the curation process are provided in the experimental section. Table 1 provides

Table 1 Overview of the dataset

CYP enzyme	Inhibitors	Non-inhibitors	Number of compounds	Total compounds	Sharing compounds
1A2	1759	1922	3681	12,369 compounds	215 compounds; 8 multi-inhibitors
2B6	84	378	462		
2C8	235	478	713		
2C9	2656	2631	5287		
2C19	1610	1674	3284		
2D6	3039	3233	6272		
3A4	5045	4218	9263		

an overview of the dataset, and the final datasets are available in Supplementary Information 1. Notably, all CYP datasets, except for CYP2B6 and CYP2C8, contained over 3,000 compounds with a balanced distribution of inhibitors and non-inhibitors. However, CYP2B6 and CYP2C8 had significantly smaller datasets (462 and 713 compounds, respectively), with a lower proportion of inhibitors. An additional challenge arises from the imbalance between high and low activity levels, particularly when selective bioactivity potency is used as a threshold for activity classification. Bioactivity potency metrics, such as IC50 is commonly employed in machine learning approaches. A higher IC50 (lower pIC50) value is preferable in the early stages of drug development to avoid DDIs. Generally, IC50 values from 1 to 40 μM are used [22, 23]; we set the inhibitor threshold to $IC50 \leq 10$ μM ($pIC50 \geq 5$), as reported by Goldwaser et al. [24], which indicates a strong inhibitor. Additionally, this threshold of $IC50 \leq 10$ μM was used to mitigate imbalanced data between inhibitors and non-inhibitors in our dataset.

This dataset encompassed seven CYP isoforms. Notably, 215 compounds were overlapped (shared) among all seven individual CYP datasets, referred to as "sharing compounds", and eight compounds were inhibitors of all seven isoforms (Table 1). Merging datasets from all seven isoforms facilitated the identification of overlapping labels across individual CYP datasets. However, this also resulted in a significant number of missing labels, particularly for the smaller CYP2B6 and CYP2C8 datasets (96% and 94% missing labels, respectively).

Visualisation of dataset

A Uniform Manifold Approximation and Projection (UMAP) plot [25] revealed that most compounds in our dataset were associated with only one CYP isoform, as shown in Fig. 1A. A small fraction (eight of the 12,369 compounds), represented by dark red dots in the plot, demonstrated multi-inhibitory activity against CYP isoforms. Interestingly, the spatial distribution of these multitarget inhibitors across the UMAP plot demonstrated a lack of clustering, implying a high degree of structural heterogeneity within this subgroup. In essence, the UMAP analysis suggested that the potent multi-inhibitory activity was not restricted to specific chemical scaffolds.

Figure 1B illustrates the chemical space distribution of CYP2B6 and CYP2C8 in comparison to the overall dataset. Both enzymes exhibited a narrow chemical space, concentrated around the central region, with a few data points separated far from the centre. CYP2B6 demonstrated a more dispersed distribution with individual data points scattered across space. In contrast, CYP2C8 exhibited a clustering pattern with multiple data points grouped together.

Construction of the baseline model

Baseline models were built using a single-task model approach for each CYP isoform using a Graph Convolutional Neural Network (GCN) algorithm [26]. The F1 and Cohens-Kappa scores were used as evaluation metrics. The averages of F1 and Cohens-Kappa ± standard deviation of the test sets are shown in Table 2.

The major CYP isoforms demonstrated robust performance in the test sets, achieving F1 scores exceeding 0.7 and kappa scores greater than 0.5. Conversely, CYP2B6 and CYP2C8 exhibited inferior performance which indicated by bold values, with F1 scores below 0.6 and kappa scores under 0.3, accompanied by larger standard deviations in the test sets. We attribute this suboptimal performance to the limited dataset size, class imbalance, and narrow structural diversity, which are factors that can hinder the development of effective predictive models. This established baseline performance served as a benchmark for comparison with subsequent models.

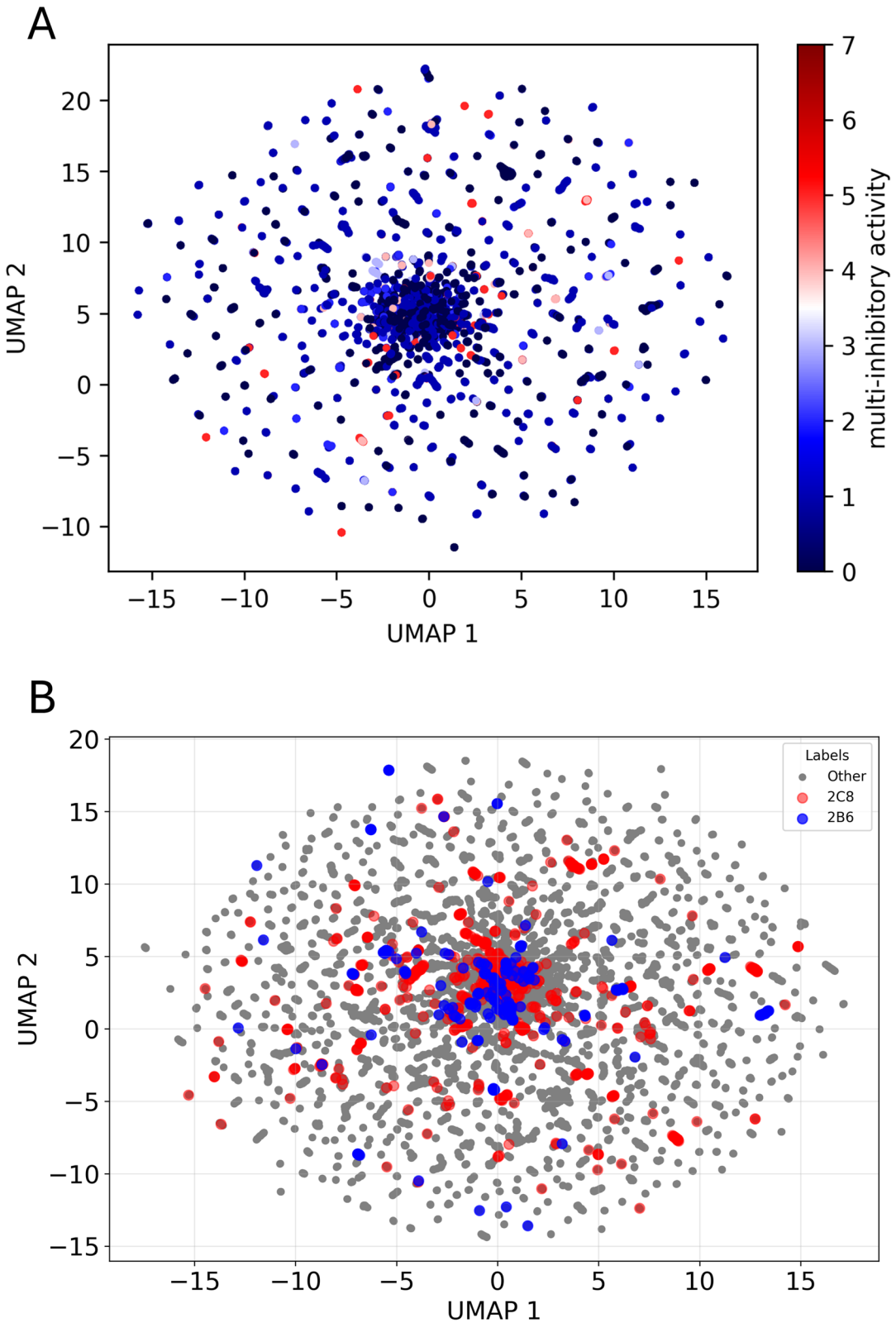


Fig. 1 The chemical space of our dataset for CYP inhibition (pIC50 ≤ 5). **A** Overall view of seven CYP isoforms, **B** CYP2B6 (blue dots) and CYP2C8 (red dots) isoforms in comparison to the overall dataset

Table 2 Evaluation of the baseline model using a single-task model

CYP isoforms	F1	SD (±)	Cohens-Kappa	SD (±)
1A2	0.811	0.031	0.646	0.052
2B6	**0.402**	0.174	**0.281**	0.198
2C8	**0.550**	0.098	**0.275**	0.136
2C9	0.796	0.015	0.578	0.028
2C19	0.775	0.041	0.564	0.066
2D6	0.819	0.036	0.650	0.053
3A4	0.844	0.009	0.645	0.020

Bold values indicate insufficient prediction performance

Construction of the multitask, fine-tuning, and multitask with imputation models for CYP2B6 and CYP2C8

To address the suboptimal performance of the baseline model for CYP2B6 and CYP2C8, we explored multitask learning and fine-tuning approaches. Multitask learning is a well-established strategy in deep learning and is known to enhance model performance by leveraging the relationships between multiple learning tasks [1, 8]. Similarly, fine-tuning has been widely adopted across various domains to improve the predictive capabilities of models by leveraging pre-trained knowledge [27–30]. We implemented both multitask learning and fine-tuning approaches using settings identical to those of the baseline model. Details of the construction of the multitask and fine-tuning models are presented in the experimental section.

Initially, the performances of the multitask and fine-tuning models based on five major CYP isoforms and multitask with imputation were compared to those of the single-task (baseline) models, as presented in Fig. 2. Overall, the multitask and fine-tuning models improved the mean F1 and Kappa scores for CYP2B6 and CYP2C8 compared to their respective baseline models. Our findings are aligned with the results of previous studies, which demonstrated that both multitask learning [1, 8] and fine-tuning [28] can improve predictive performance. However, we failed to achieve statistically significant improvements over the baseline models. We suspect that this limited improvement was attributed to class imbalance in the CYP2B6 dataset, with a predominance of non-inhibitors, which may have contributed to model instability and increased the variance in performance across all model types (single-task, multitask, and fine-tuning). As reported by Li et al. [31], a class imbalance can lead to biased models, favouring the majority class and decreasing the predictive performance of the minority class. Also, the insignificant improvement was probably associated with the training setup, in which missing values were assigned a weight of zero and were ignored during the evaluation. This approach may introduce bias towards non-inhibitor predictions.

To achieve statistically significant improvements, we first analysed the effects of missing data and sharing compounds on the multitask model. Because merging multiple tasks results in the creation of missing labels/missing values (blank entries) for untested compounds, we conducted an additional study on the significance of data sharing and missing labels on multitask learning performance. We systematically varied the percentage of missing labels in one dataset with no missing labels, ranging from 0 to 95%. While the kMol platform treats missing labels (blank entries) as ignored data points, the model performance, measured by F1 and Kappa scores, exhibited a significant decline when the proportion of missing data exceeded 50% (Figure S1 in Supplementary Information 2). This finding aligns with previous observations by Ayilara et al. [31], who reported the detrimental effects of missing data on model performance, data analysis accuracy, and the potential for biased outcomes, particularly within clinical registry datasets. We found that fewer missing labels and larger sharing compounds resulted in better performance.

Given those observations, we decided to implement imputation on the missing value (missing label) for our training set using the predicted label. We employed a multi-imputation strategy using single-task, fine-tuning, and multitask model prediction results to mask the missing labels (blank data) in our datasets [32–35]. This approach leverages information from the observed data to estimate missing entries. Subsequently, we incorporated the imputed datasets into our multitask models to predict CYP activity in small datasets, as illustrated in Scheme 1 in the experimental section. The final evaluation was performed using the test sets of CYP2B6 and CYP2C8, which were held at the beginning of the analysis and were not included in the training set.

Our multitask models, in combination with data imputation of missing values by predicted labels from the single-task (MIPS), multitask (MIPM), and fine-tuning (MIPFT) models, exhibited even greater improvement over the baseline model's prediction performance for CYP2B6 and CYP2C8 (the three bars on the right at the plots in Fig. 2). In particular, MIPM exhibited a statistically significant improvement in CYP2B6 ($p < 0.05$) and CYP2C8 ($p < 0.01$) inhibitor prediction. In addition, MIPS also demonstrated a significant improvement ($p < 0.01$) in CYP2C8 inhibitor prediction.

By imputing missing values with predicted labels, we provided a model with more accurate training data, leading to improved predictive performance. This result

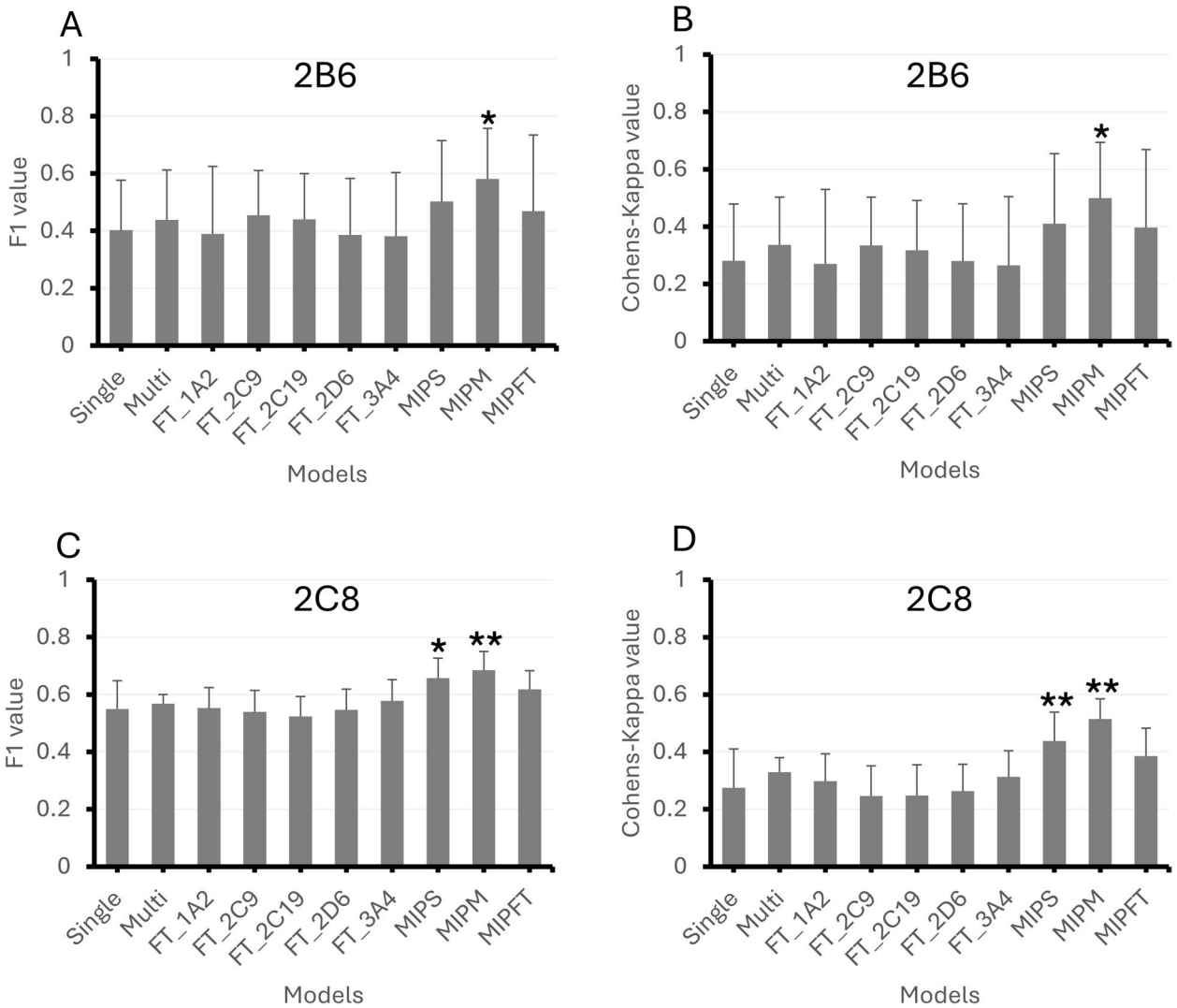
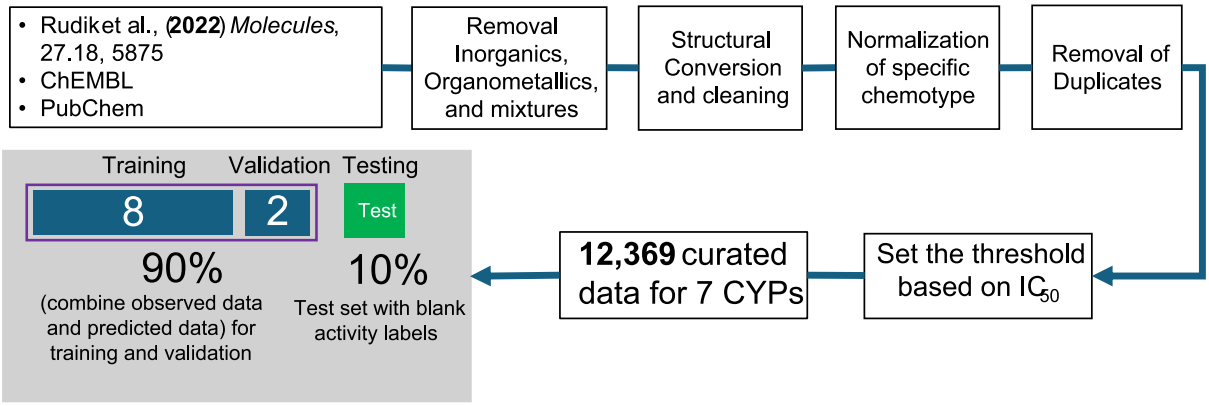


Fig. 2 Comparison of model performance for CYP2B6 (**A** F1, **B** Cohens-Kappa) and CYPC8 (**C** F1, **D** Cohens-Kappa). Single: single-task learning/baseline; Multi: multitask learning based on CYP isoform data; FT_CYP isoform: fine-tuning models based on CYP isoforms; MIPS: multitask learning with the imputation of predicted label from single-task model; MIPM: multitask learning with the imputation of predicted label from multitask model; MIPFT: multitask learning with the imputation of predicted label from fine-tuning model. *Significant compared to single-task model ($p < 0.05$), **Significant compared to single-task model ($p < 0.01$)

aligns with a previous study conducted by Hasan et al. [36], who removed outliers and filled in missing values, providing improvements in their diabetes prediction model.

Our multitask model with the imputation approach was particularly effective for CYP2C8, likely because of its larger dataset and balanced class distribution compared to CYP2B6. The prediction results for CYP2C8 were more accurate than the single-task model (Kappa = 0.275 ± 0.136) with representative Kappa scores of 0.438 ± 0.101, 0.629 ± 0.222, and 0.385 ± 0.098, respectively, for MIPS, MIPM, and MIPFT. In addition, we suspect that our multitask models with imputation were able to capture overall trends in the data, particularly when using multitask learning. In CYP2B6 prediction, only MIPM exhibited statistically significant improvement (Kappa = 0.514 ± 0.071) compared to the baseline model (Kappa = 0.281 ± 0.198). We also observed inconsistent prediction performance on CYP2B6, indicated by larger standard deviations, suggesting that the imputed values may not have accurately captured the underlying trends, potentially leading to higher false negative rates.



Scheme 1 Performing missing labels (blank) prediction and multitask with imputation

CYP inhibition potential on approved drugs for CYP2B6 and CYP2C8

Building on our models with imputed data, we aimed to predict the inhibitory activity of approved drugs against two specific CYP isoforms, CYP2B6 and CYP2C8, which have small datasets and are well-mentioned in the FDA guidelines [16]. Moreover, many compounds have been untested against CYP2B6 or CYP2C8 inhibition. To address these issues, we created a dataset of 1,808 human-approved drugs (available in Supplementary Information 1, Sheet 2) taken from DrugBank [37], excluding those included in our dataset, defined as dataset_approved_drugs. Dataset_approved_drugs contained 26 and 55 approved drugs that were reported as CYP2B6 and CYP2C8 inhibitors, respectively. The remaining drugs lacked known inhibitory information for these isoforms. We used these 26 and 55 known inhibitors as external test sets for evaluation. The external test set was subjected to preprocessing procedures identical to those employed for the training and validation sets. Recall quantification was employed to determine the model reliability of known approved drugs as an external test set for CYP2B6 and CYP2C8 using the MIPS, MIPM, and MIPFT models. We also constructed ensemble models that derived conclusions based on majority voting of the results of these three models.

As detailed in Table 3, the MIPS model achieved the highest recall scores, reaching 0.27 for CYP2B6 and MIPM reaching 0.60 for CYP2C8. This result suggests that inhibitors of CYP2C8 were successfully predicted in the external test set but not for CYP2B6. One reason for the inadequate prediction accuracy of CYP2B6 was the structural similarity between approved_drugs dataset and our dataset. Comparison of chemical spaces based on structural information showed that, while occupying similar overall chemical spaces, the Tanimoto coefficients were relatively low, with a mean of 0.39 and ranging from 0.05 to 1.0 (Supplementary Information 2, Figure S3). Consequently, our model may struggle to accurately identify compounds with these divergent structures, resulting in a low recall score. Interestingly, both the MIPM and MIPFT models exhibited comparable performance in predicting CYP2C8 inhibitors, and even our ensemble model with majority voting decisions did not surpass the MIPM recall value, suggesting that the agreement among the three models may not have been strong enough to improve the prediction accuracy. Although the performance of CYP2B6 remains insufficient, we

Table 3 Prediction performance on observed data of CYP inhibitor drugs

Model	2B6			2C8		
	TP	FN	Recall	TP	FN	Recall
MIPS	**7**	**19**	**0.27**	24	31	0.44
MIPM	5	21	0.19	**33**	**22**	**0.60**
MIPFT	2	24	0.07	32	23	0.58
Ensemble model	5	21	0.19	32	23	0.58

Bold values indicate insufficient prediction performance

MIPS multitask learning with imputation of the predicted label from the single-task model, *MIPM* multitask learning with imputation of the predicted label from the multitask model, *MIPFT* multitask learning with imputation of the predicted label from the fine-tuning model. *TP* true positive, *FN* false negative

hypothesized that incorporating structural information, specifically maximum Tanimoto similarity, would improve the ranking of potential inhibitors, particularly for CYP2B6. The correlation analysis between maximum Tanimoto similarity and F1 score, detailed in Supplementary Information 2 Figure S2, revealed that test compounds with higher similarity to the training compounds would yield more accurate prediction results.

We then applied our procedure to the remaining drugs in dataset_approved_drugs. In CYP2B6 prediction, we used the best model to predict drug inhibitors of CYP2B6. In contrast, for CYP2C8 prediction, we identified potential inhibitors by all three models to obtain more reliable prediction results (Table 4). Subsequently, our prediction models identified 161 and 154 candidates for CYP2B6 and CYP2C8, respectively (Supplementary Information 1, Sheets 3 and 4). Additionally, our multitask learning approach, which leveraged information across all seven CYP isoforms, facilitated the prediction of compounds capable of inhibiting several CYP isoforms. Our models identified 30 approved drugs potentially inhibiting both CYP2B6 and CYP2C8, including two compounds predicted to inhibit all seven CYP isoforms (Table S1, Supplementary Information 2). Thus, in future work, we plan to utilise this prediction knowledge to build comprehensive drug-drug interaction predictions.

To mitigate potential mispredictions, a composite scoring metric was implemented, integrating a probability score (weighted 0.7) and the maximum Tanimoto similarity (weighted 0.3). The probability score reflects the likelihood produced by the model, while the maximum Tanimoto similarity quantifies how reliable the model itself would be when applied to a given chemical structure. This composite score was then used to rank potential inhibitors, thereby reducing the likelihood of misclassification.

The top 10 potential inhibitors ranked by the composite score are presented in Table 5. Simeprevir and Lercanidipine were identified as the top drug candidates for CYP2B6 and CYP2C8, respectively. Simeprevir is an antiviral agent that inhibits HCV NS3/4A protease to treat chronic hepatitis C virus (HCV) and primarily metabolized by CYP3A [38]. Moreover, Lercadipine is an anti-hypertension drug that belongs to a class of calcium channel blockers. Lercadipine is well-known metabolized by CYP3A4 and potentially inhibits CYP2D6 and CYP3A4 [39]. Notably, no documented evidence of Lercadipine exhibiting CYP2B6 and CYP2C8 inhibitory activity have been found in existing repositories.

Conclusion

Our study emphasises the significant challenges posed by limited data, such as small sample sizes, regarding the efficacy of single-task prediction models for CYP inhibition. We successfully constructed prediction models for CYP2B6 and CYP2C8 using multitask deep learning, particularly multitask learning with a graph convolutional network (GCN), to overcome the limitations associated with small datasets. Compared with single-task models, fine-tuning and multitask models have resulted in substantial improvements. Ultimately, the most effective strategy for accurate CYP inhibition prediction involved multitask models that incorporated data imputation techniques and surpassed all other models in predicting the CYP inhibitors of CYP2B6 and CYP2C8 using the observed data. In addition, this approach successfully predicted the CYP inhibitory activities of 1,808 approved drugs. We also highlighted 161 and 154 potential inhibitors of CYP2B6 and CYP2C8, respectively, which warrant further experimental validation.

Due to the limited amount of available data, it is often difficult to build a predictive model for pharmacokinetic parameters. This study demonstrates that prediction models can be effectively constructed even with small datasets by utilising related data extensively. Furthermore, beyond identifying individual CYP-drug interactions, this study paves the way for the indirect discovery of multi-inhibitor drugs.

Methods/experimental

Dataset, data curation, and molecular representation

Chemical and activity data retrieved from the PubChem Bioassay Database (consisted of AID410-1A2—9174 data; AID883-2C9—10,296 data; AID899-2C19—10,296 data; AID891-2D6—10,296 data; AID885-3A4—14,115

Table 4 Overall prediction for CYP2B6 and CYP2C8 using multitask learning with imputation

Model	2B6		2C8		Final potential inhibitor		Inhibitor
	inh	non-inh	inh	non-inh	2B6-inh	2C8-inh	2B6-2C8
MIPS	161	1621	527	1226			
MIPM	–	–	411	1342	161	154	30
MIPFT	–	–	388	1365			

Table 5 Top potential CYP2B6 and CYP2C8 inhibitor drugs based on the highest composite score

CYP2B6 potential inhibitors						
No.	**DrugBank ID**	**CHEMBL ID**	**Drug's name**	**Probability score**	**Maximum Tanimoto's similarity**	**Composite score (0.7 prob + 0.3 max. Tanimoto's sim.)**
1	DB06290	CHEMBL501849	Simeprevir	0.979	0.928	0.964
2	DB11633	CHEMBL409153	Isavuconazole	1.000	0.868	0.960
3	DB00377	CHEMBL1189679	Palonosetron	0.997	0.792	0.935
4	DB11340	–	Ubiquinol	1.000	0.780	0.934
5	DB00758	CHEMBL1771	Clopidogrel	0.984	0.771	0.920
6	DB11254	CHEMBL443605	Hexylresorcinol	0.999	0.730	0.919
7	DB14120	CHEMBL3961037	Phenylethyl resorcinol	0.999	0.727	0.918
8	DB00735	CHEMBL626	Naftifine	0.992	0.724	0.912
9	DB05239	CHEMBL2146883	Cobimetinib	1.000	0.700	0.910
10	DB12612	CHEMBL3707247	Ozanimod	1.000	0.694	0.964
CYP2C8 potential inhibitors						
No.	**DrugBank ID**	**CHEMBL ID**	**Drug's name**	**Probability score**	**Maximum Tanimoto's similarity**	**Composite score (0.7 prob + 0.3 max. Tanimoto's sim.)**
1	DB00528	CHEMBL250270	Lercanidipine	0.998	0.912	0.972
2	DB09238	CHEMBL1085699	Manidipine	0.988	0.910	0.965
3	DB14086	CHEMBL311498	Cianidanol	0.971	0.875	0.942
4	DB13946	CHEMBL2107067	Testosterone undecanoate	1.000	0.783	0.935
5	DB11340	–	Ubiquinol	1.000	0.780	0.934
6	DB13944	CHEMBL1200335	Testosterone enanthate	0.999	0.783	0.934
7	DB14989	CHEMBL3948730	Umbralisib	1.000	0.746	0.924
8	DB13943	CHEMBL1201101	Testosterone cypionate	1.000	0.739	0.922
9	DB14914	CHEMBL3545253	Flortaucipir F-18	0.999	0.727	0.918
10	DB12364	CHEMBL512351	Betrixaban	0.999	0.714	0.972

data, AID884-3A4—14,115 data), the ChEMBL Database (consisted of 1A2—4812 data; 2B6—572 data; 2C8—795 data; 2C9—6366 data; 2C19—4303 data; 2D6—7231 data; 3A4—11,712 data), and Rudik et al. [21] yielding a total of 170,355 data points for 1A2, 2B6, 2C8, 2C9, 2C19, 2D6, and 3A4.

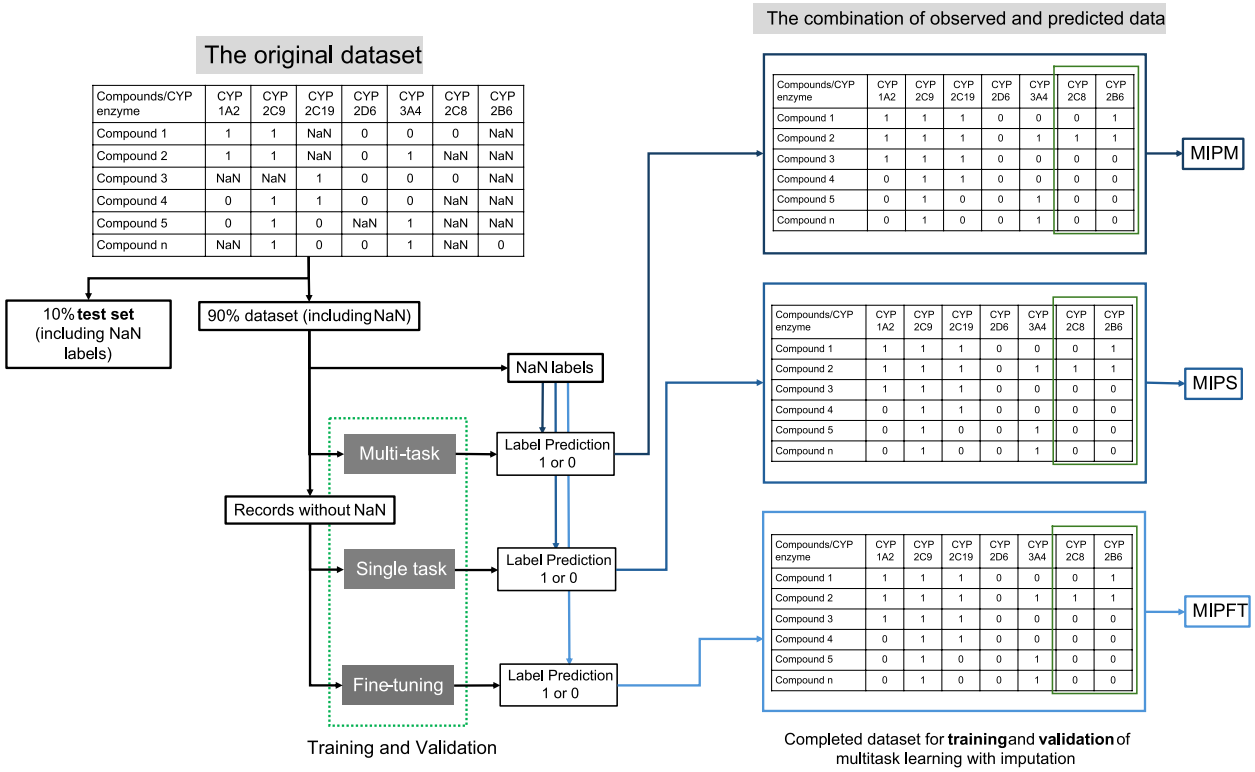
Data curation was performed using the KNIME [40] analytics platform (version 4.7.2) equipped with OpenBabel and RDKit libraries. Additionally, Python 3 with the Pandas, RDKit, and MolVS modules were employed for this process. This process yielded a final dataset of 12,369 compounds tested for the inhibition of seven CYP isoforms. For compounds with multiple data points, the lowest IC50 value was selected based on experimental conditions, such as concentration and incubation time, as these factors can influence inhibitory potency. Classification was based on the following criteria: IC50 ≤ 10 μM was classified as an inhibitor, IC50 > 10 μM as a non-inhibitor, and unclear data points with less than values that more than 10 μM (e.g., IC50 < 100 μM) and greater than values that less than 10 μM (e.g., IC50 > 1 μM) were excluded from our dataset. The workflow for data curation is illustrated in Scheme 2.

Data visualisation

We generated a UMAP plot to visualise the relationships between compounds based on their fingerprint features derived from the SMILES strings using Morgan Fingerprint, which generates a 1024-bit fingerprint for each compound. A UMAP plot was generated using the UMAP package in Python 3 [25].

Model construction and optimisation

We employed a graph convolutional network (GCN) implemented in the kMoL library (version 1.1.5) [26] for model construction. The kMoL packages can be cloned from https://github.com/elix-tech/kmol.git. The kMoL is a specialised library designed to build machine learning models applicable to drug discovery and life science research. Our model approach employed kMoL. The kMoL platform proceeds with SMILES input using



Scheme 2 Data curation workflow

graph featurisation. It then generates atomic features and an adjacency matrix before being fed to several graph convolutional layers. The final layer is propagated using global max pooling and global add pooling. The final prediction was generated as the output.

We prepared 10 different data batches with 10 different random split seeds. 10% of the datasets were used as the test sets. The remaining 90% of the data were divided into training and validation sets at a ratio of 8:2 (as shown in Scheme 2). Model optimisation was achieved using a validation set of over 200 epochs with fivefold nested cross-validation. Hyperparameters, such as hidden features, dropout rate, layer type and number, residuals, and batch size, were optimised using Optuna (100 trials). The F1 and Kappa scores served as primary evaluation metrics. The final models were subsequently evaluated using the test set.

The baseline (single-task) model construction

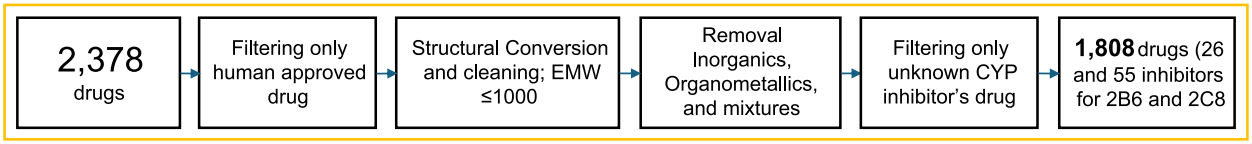
A baseline or single-task model was constructed for each CYP isoform (CYP1A2, CYP2B6, CYP2C8, CYP2C9, CYP2C19, CYP2D6, and CYP3A4) using the corresponding compound dataset as input. The average F1 and Kappa scores and their standard deviations were calculated for each model. The configuration file and the example invocations for the single-task model is provided in Supplementary Information 3.

The multitask model construction

Multitask deep learning was implemented by combining all seven CYP isoforms (CYP1A2, CYP2B6, CYP2C8, CYP2C9, CYP2C19, CYP2D6, and CYP3A4) as the multitask input. The average F1 and Kappa scores and their standard deviations were calculated for each multitask model. The configuration file and the example invocations for the multitask learning model is provided in Supplementary Information 3.

The fine-tuning model construction

The fine-tuning models were retrieved from kMol documentation. The document path of the trained model was placed in the *checkpoint_path* in the *config.json* file. Additionally, we set the "*is_finetuning*" to true at *config.json*. The new dataset was assigned to the "*input_path*" in *config.json*. The example invocations for the multitask learning model is provided in Supplementary Information 3.



Scheme 3 Curation process of approved drugs for the external test set

Performing missing labels (blank) prediction and multitask model with predicted label imputation

Initially, the original dataset was randomly split into 90% and 10%. To predict the CYP-inhibitory activities of compounds lacking known inhibitory labels against seven CYP isoforms, we generated and extracted all the missing labels (missing values, NaNs) from the original dataset. We then predicted the missing (NaN) labels using the best model from each single-task, multitask, and fine-tuning model. Thus, we merged the predicted results to mask the missing (NaN) labels in the training and validation sets, and the completed datasets were established. The remaining 10% of the original dataset, which contained compounds with missing activity labels for certain CYP isoforms, was used as the test set to evaluate the model. Average F1 and Kappa scores along with their standard deviations were calculated for each evaluation. We prepared 10 different data batches with 10 different random split seeds.

Prediction of potential CYP2B6 and CYP2C8 inhibitors among available approved drugs

A total of 2378 approved drugs (small molecules with a molecular weight less than or equal to 1 kDa) were retrieved from the DrugBank database [41] using the KNIME [40] analytics platform (version 4.7.2). We filtered drugs that were not included in our training-validation test set, resulting in 1,808 unique drugs (Scheme 3). To evaluate model reliability, we quantified the recall for known inhibitors of CYP2B6 and CYP2C8 (26 and 55 inhibitors, respectively), which were set as the external test set. The final model with the best performance, identified by kMol, was employed to predict the inhibitory activity of the approved drugs against CYP isoforms.

To ascertain the most promising inhibitor for CYP2B6 and CYP2C8, a composite scoring function was employed. This function integrated a probability score and a Tanimoto similarity score, weighted 7:3, respectively. The ranking of potential inhibitors was determined based on the highest composite score. This weighting was chosen to prioritise the probability score as the main factor and, as additional information, we included structural similarity to the compounds in our training set (measured by the Tanimoto coefficient), to provide an additional layer of information about model reliability. We also explored alternative ratios (6:4, 8:2, and 9:1). The top-ranked compounds remained consistent across different weight ratios, although some compounds in the middle of the ranking changed positions by one or two places, reflecting small differences in their weighted scores. This ranking system will be used to prioritize compounds for further experimental validation, allowing us to focus on the most promising potential inhibitors. In more detail, a probability score was derived from a sigmoid transformation of kMol-generated logits, and Tanimoto similarity was calculated by comparing the chemical structures represented using 167-bit MACCS fingerprints between the predicted drugs list and known inhibitors within the original CYP2B6 and CYP2C8 datasets.

Statistical analysis

We used a simple t-test analysis for two independent samples to evaluate the significance ($p < 0.05$ and $p < 0.01$) between all models compared to the single-task model for improving the CYP inhibitor prediction for a small dataset (CYP2B6 and CYP2C8).

Abbreviations

CYP	Cytochrome P450
QSAR	Quantitative structure–activity relationship
SMILES	Simplified molecular-input line-entry system
GCN	Graph convolutional network
FT	Fine-tuning
MIPS	Multitask learning with imputation of predicted label from single-task model
MIPM	Multitask learning with imputation of predicted label from multitask model
MIPFT	Multitask learning with imputation of predicted label from fine-tuning model
TP	True positive
FN	False negative
EMW	Exact molecular weight
FDA	Food and Drug Administration of USA

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13321-025-01015-2.

Supplementary Material 1.

Supplementary Material 2.

Acknowledgements
This research was conducted in "Development of a Next-generation Drug Discovery AI through Industry-academia Collaboration (DAIIA)" supported by the Japan Agency for Medical Research and Development (AMED) under Grant Number JP22nk0101111. This study was in part supported by OU Master Plan Implementation Project. We would like to thank Editage (http://www.editage.com) for editing and reviewing this manuscript for English language.

Author contributions
E.E.P.: Framework design, data collection, data curation, investigation, formal analysis, validation, visualisation, writing—original draft, writing—review, and editing. R.W.: methodology, formal analysis, writing, review, editing, and supervision. K.M.: conceptualisation, writing, review, editing, and supervision.

Funding
This research was conducted in "Development of a Next-generation Drug Discovery AI through Industry-academia Collaboration (DAIIA)" supported by Japan Agency for Medical Research and Development (AMED) under Grant Number JP22nk0101111. This study was in part supported by OU Master Plan Implementation Project.

Availability of data and materials
All the data is provided within the manuscript or supplementary information files.

Declarations

Competing interests
The authors declare no competing interests.

Received: 7 January 2025 Accepted: 13 April 2025
Published online: 30 April 2025

References

1. Nguyen-Vo T-H, Trinh QH, Nguyen L, Nguyen-Hoang P-U, Nguyen T-N, Nguyen DT, Nguyen BP, Le L (2022) iCYP-MFE: identifying human cytochrome P450 inhibitors using multitask learning and molecular fingerprint-embedded encoding. J Chem Inf Model 62:5059–5068. https://doi.org/10.1021/acs.jcim.1c00628
2. Ogu CC, Maxa JL (2000) Drug interactions due to cytochrome P450. Bayl Univ Med Cent Proc 13:421–423. https://doi.org/10.1080/08998280.2000.11927719
3. Esteves F, Rueff J, Kranendonk M (2021) The central role of cytochrome P450 in xenobiotic metabolism—a brief review on a fascinating enzyme family. J Xenobiot 11:94–114. https://doi.org/10.3390/jox11030007
4. Li L, Lu Z, Liu G, Tang Y, Li W (2023) Machine Learning Models to Predict Cytochrome P450 2B6 Inhibitors and Substrates. Chem Res Toxicol 36:1332–1344. https://doi.org/10.1021/acs.chemrestox.3c00065
5. Tian S, Djoumbou-Feunang Y, Greiner R, Wishart DS (2018) CypReact: a software tool for in silico reactant prediction for human cytochrome P450 enzymes. J Chem Inf Model 58:1282–1291. https://doi.org/10.1021/acs.jcim.8b00035
6. Deodhar M, Al Rihani SB, Arwood MJ, Darakjian L, Dow P, Turgeon J, Michaud V (2020) Mechanisms of CYP450 inhibition: understanding drug-drug interactions due to mechanism-based inhibition in clinical practice. Pharmaceutics 12:846. https://doi.org/10.3390/pharmaceutics12090846
7. Li X, Xu Y, Lai L, Pei J (2018) Prediction of human cytochrome P450 inhibition using a multitask deep autoencoder neural network. Mol Pharm 15:4336–4345. https://doi.org/10.1021/acs.molpharmaceut.8b00110
8. Ai D, Cai H, Wei J, Zhao D, Chen Y, Wang L (2023) DEEPCYPs: A deep learning platform for enhanced cytochrome P450 activity prediction. Front Pharmacol 14:1099093. https://doi.org/10.3389/fphar.2023.1099093
9. Banerjee P, Dunkel M, Kemmler E, Preissner R (2020) SuperCYPsPred—a web server for the prediction of cytochrome activity. Nucleic Acids Res 48:W580–W585. https://doi.org/10.1093/nar/gkaa166
10. Cheng F, Yu Y, Shen J, Yang L, Li W, Liu G, Lee PW, Tang Y (2011) Classification of cytochrome P450 inhibitors and noninhibitors using combined classifiers. J Chem Inf Model 51:996–1011. https://doi.org/10.1021/ci200028n
11. Plonka W, Stork C, Šícho M, Kirchmair J (2021) CYPlebrity: machine learning models for the prediction of inhibitors of cytochrome P450 enzymes. Bioorg Med Chem 46:116388. https://doi.org/10.1016/j.bmc.2021.116388
12. Zhang T, Dai H, Liu LA, Lewis DFV, Wei D (2012) Classification models for predicting cytochrome P450 enzyme-substrate selectivity. Mol Inform 31:53–62. https://doi.org/10.1002/minf.201100052
13. Sun H, Veith H, Xia M, Austin CP, Huang R (2011) Predictive models for cytochrome P450 isozymes based on quantitative high throughput screening data. J Chem Inf Model 51:2474–2481. https://doi.org/10.1021/ci200311w
14. Raunio H, Kuusisto M, Juvonen RO, Pentikäinen OT (2015) Modeling of interactions between xenobiotics and cytochrome P450 (CYP) enzymes. Front Pharmacol. https://doi.org/10.3389/fphar.2015.00123
15. Backman JT, Filppula AM, Niemi M, Neuvonen PJ (2016) Role of cytochrome P450 2C8 in drug metabolism and interactions. Pharmacol Rev 68:168–241. https://doi.org/10.1124/pr.115.011411
16. U.S. Food and Drug Administration (2012) Guidance for industry: drug interaction studies - study design, data analysis, implications for dosing, and labeling recommendations
17. U.S. Food and Drug Administration (2006) Guidance for industry: drug interaction studies - study design, data analysis, and implications for dosing and labeling
18. Xu P, Ji X, Li M, Lu W (2023) Small data machine learning in materials science. NPJ Comput Mater 9:42. https://doi.org/10.1038/s41524-023-01000-z
19. Mendez D, Gaulton A, Bento AP, Chambers J, De Veij M, Félix E, Magariños MP, Mosquera JF, Mutowo P, Nowotka M, Gordillo-Marañón M, Hunter F, Junco L, Mugumbate G, Rodriguez-Lopez M, Atkinson F, Bosc N, Radoux CJ, Segura-Cabrera A, Hersey A, Leach AR (2019) ChEMBL: towards direct deposition of bioassay data. Nucleic Acids Res 47:D930–D940. https://doi.org/10.1093/nar/gky1075
20. Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, Li Q, Shoemaker BA, Thiessen PA, Yu B, Zaslavsky L, Zhang J, Bolton EE (2023) PubChem 2023 update. Nucleic Acids Res 51:D1373–D1380. https://doi.org/10.1093/nar/gkac956
21. Rudik A, Dmitriev A, Lagunin A, Filimonov D, Poroikov V (2022) Computational prediction of inhibitors and inducers of the major isoforms of cytochrome P450. Molecules 27:5875. https://doi.org/10.3390/molecules27185875
22. Konda LSK, Keerthi Praba S, Kristam R (2019) hERG liability classification models using machine learning techniques. Comput Toxicol 12:100089. https://doi.org/10.1016/j.comtox.2019.100089
23. Zhang X, Zhao P, Wang Z, Xu X, Liu G, Tang Y, Li W (2021) In silico prediction of CYP2C8 inhibition with machine-learning methods. Chem Res Toxicol 34:1850–1859. https://doi.org/10.1021/acs.chemrestox.1c00078
24. Goldwaser E, Laurent C, Lagarde N, Fabrega S, Nay L, Villoutreix BO, Jelsch C, Nicot AB, Loriot M-A, Miteva MA (2022) Machine learning-driven identification of drugs inhibiting cytochrome P450 2C9. PLOS Comput Biol 18:e1009820. https://doi.org/10.1371/journal.pcbi.1009820
25. McInnes L, Healy J, Saul N, Großberger L (2018) UMAP: uniform manifold approximation and projection. J Open Source Softw 3(29):861. https://doi.org/10.21105/joss.00861
26. Kojima R, Ishida S, Ohta M, Iwata H, Honma T, Okuno Y (2020) kGCN: a graph-based deep learning framework for chemical structures. J Cheminform 12:32. https://doi.org/10.1186/s13321-020-00435-6
27. Öztürk C, Taşyürek M, Türkdamar MU (2023) Transfer learning and fine-tuned transfer learning methods' effectiveness analyse in the CNN-based deep learning models. Concurr Comput Pract Exp 35:e7542. https://doi.org/10.1002/cpe.7542
28. He Z, Zhang L, Wang H (2023) An initial prediction and fine-tuning model based on improving GCN for 3D human motion prediction. Front

Comput Neurosci 17:1145209. https://doi.org/10.3389/fncom.2023.1145209
29. Abbas A, Shah AN, Tanveer M, Ahmed W, Shah AA, Fiaz S, Waqas MM, Ullah S (2022) MiRNA fine tuning for crop improvement: using advance computational models and biotechnological tools. Mol Biol Rep 49:5437–5450. https://doi.org/10.1007/s11033-022-07231-5
30. Bohmrah MK, Kaur H (2021) Classification of Covid-19 patients using efficient fine-tuned deep learning DenseNet model. Glob Transit Proc 2:476–483. https://doi.org/10.1016/j.gltp.2021.08.003
31. Ayilara OF, Zhang L, Sajobi TT, Sawatzky R, Bohm E, Lix LM (2019) Impact of missing data on bias and precision when estimating change in patient-reported outcomes from a clinical registry. Health Qual Life Outcomes 17:106. https://doi.org/10.1186/s12955-019-1181-2
32. Emmanuel T, Maupong T, Mpoeleng D, Semong T, Mphago B, Tabona O (2021) A survey on missing data in machine learning. J Big Data 8:140. https://doi.org/10.1186/s40537-021-00516-9
33. Donders ART, Van Der Heijden GJMG, Stijnen T, Moons KGM (2006) Review: a gentle introduction to imputation of missing values. J Clin Epidemiol 59:1087–1091. https://doi.org/10.1016/j.jclinepi.2006.01.014
34. Shadbahr T, Roberts M, Stanczuk J, Gilbey J, Teare P, Dittmer S, Thorpe M, Torné RV, Sala E, Lió P, Patel M, Preller J, Collaboration AIX-COVNET, Selby I, Breger A, Weir-McCall JR, Gkrania-Klotsas E, Korhonen A, Jefferson E, Langs G, Yang G, Prosch H, Babar J, Escudero Sánchez L, Wassin M, Holzer M, Walton N, Lió P, Rudd JHF, Mirtti T, Rannikko AS, Aston JAD, Tang J, Schönlieb C-B (2023) The impact of imputation quality on machine learning classifiers for datasets with missing values. Commun Med 3:139. https://doi.org/10.1038/s43856-023-00356-z
35. Hasan MdK, Alam MdA, Roy S, Dutta A, Jawad MdT, Das S (2021) Missing value imputation affects the performance of machine learning: a review and analysis of the literature (2010–2021). Inform Med Unlocked 27:100799. https://doi.org/10.1016/j.imu.2021.100799
36. Hasan MdK, Alam MdA, Das D, Hossain E, Hasan M (2020) Diabetes prediction using ensembling of different machine learning classifiers. IEEE Access 8:76516–76531. https://doi.org/10.1109/ACCESS.2020.2989857
37. Knox C, Wilson M, Klinger CM, Franklin M, Oler E, Wilson A, Pon A, Cox J, Chin NE, Strawbridge SA, Garcia-Patino M, Kruger R, Sivakumaran A, Sanford S, Doshi R, Khetarpal N, Fatokun O, Doucet D, Zubkowski A, Rayat DY, Jackson H, Harford K, Anjum A, Zakir M, Wang F, Tian S, Lee B, Liigand J, Peters H, Wang RQ, Nguyen T, So D, Sharp M, da Silva R, Gabriel C, Scantlebury J, Jasinski M, Ackerman D, Jewison T, Sajed T, Gautam V, Wishart DS (2024) DrugBank 6.0: the DrugBank Knowledgebase for 2024. Nucleic Acids Res 52:D1265–D1275. https://doi.org/10.1093/nar/gkad976
38. Ouwerkerk-Mahadevan S, Snoeys J, Peeters M, Beumont-Mauviel M, Simion A (2016) Drug-drug interactions with the NS3/4A protease inhibitor Simeprevir. Clin Pharmacokinet 55:197–208. https://doi.org/10.1007/s40262-015-0314-y
39. Ferri N, Corsini A, Pontremoli R (2024) Antihypertensive and renal protection effects of lercanidipine and lercanidipine/enalapril: Renal protection by lercanidipine. Eur Atheroscler J 3:73–80. https://doi.org/10.56095/eaj.v3i3.58
40. Berthold MR, Cebron N, Dill F, Gabriel TR, Kötter T, Meinl T, Ohl P, Sieb C, Thiel K, Wiswedel B (2007) KNIME: The Konstanz Information Miner. In: Studies in classification, data analysis, and knowledge organization (GfKL 2007). Springer
41. Knox C, Law V, Jewison T, Liu P, Ly S, Frolkis A, Pon A, Banco K, Mak C, Neveu V, Djoumbou Y, Eisner R, Guo AC, Wishart DS (2011) DrugBank 3.0: a comprehensive resource for "Omics" research on drugs. Nucleic Acids Res 39:D1035–D1041. https://doi.org/10.1093/nar/gkq1126

Publisher's Note