**An uncertainty-guided deep learning method facilitates rapid screening of CYP3A4 inhibitors**

Ruixuan Wang§, Zhikang Liu‡, Jiahao Gong§, Qingping Zhou\*‡, Xiaoqing Guan\*§, Guangbo Ge\*§.

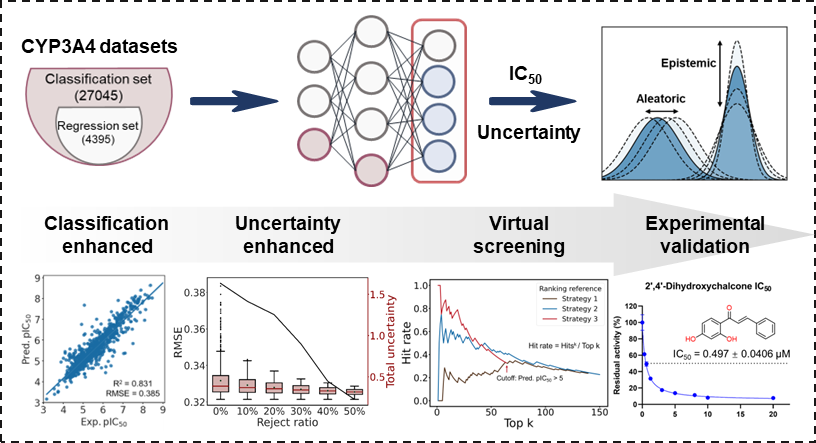
§Shanghai Frontiers Science Center of TCM Chemical Biology, Institute of Interdisciplinary Integrative Medicine Research, Shanghai University of Traditional Chinese Medicine, Shanghai 201203, China.

**‡**Central South University, School of Mathematics and Statistics, Changsha 410083, China.

KEY WORDS:CYP3A4, regression model, uncertainty, virtual screening

**ABSTRACT**

Cytochrome P450 3A4 (CYP3A4), a prominent member of the P450 enzyme superfamily, plays a crucial role in metabolizing various xenobiotics, including over 50% of clinically significant drugs. Evaluating CYP3A4 inhibition before drug approval is essential to avoid potentially harmful pharmacokinetic drug-drug interactions (DDIs) and adverse drug reactions (ADRs). Despite the development of several CYP inhibitor prediction models, the primary approach for screening CYP inhibitors still rely on experimental methods. This might stem from the limitations of existing models, which only provide deterministic classification outcomes instead of precise inhibition intensity (e.g., IC50), and often suffer from inadequate prediction reliability. To address this challenge, we propose an uncertainty-guided regression model to accurately predict IC50 values of anti-CYP3A4 activities. First, a comprehensive dataset of CYP3A4 inhibitors was compiled, consisting of 27,045 compounds with classification labels, including 4,395 compounds with explicit IC50 values. Second, by integrating the predictions of the classification model trained on a larger dataset and introducing an evidential uncertainty method to rank prediction confidence, we obtained a high-precision and reliable regression model. Finally, we use the evidential uncertainty values as a trustworthy indicator to perform virtual screening of an in-house compound set. The *in vitro* experiments results revealed that this new indicator significantly improved the hit ratio and reduced false positives among the top-ranked compounds. Specifically, among the top 20 compounds ranked with uncertainty, 15 compounds were identified as novel CYP3A4 inhibitors, and three of them exhibited activities less than 1 μM. In summary, our findings highlight the effectiveness of incorporating uncertainty in compound screening, providing a promising strategy for drug discovery and development.

****

# 1. INTRODUCTION

The mammalian cytochrome P450 (CYP) superfamily is an important class of heme-containing enzymes, which play crucial roles in oxidative metabolism of a wide range of endogenous substances and xenobiotics. In humans, a total of 57 CYP isoforms have been identified, which have been classified into 44 subfamilies based on their amino acid identities.1,2 Among all identified human CYP isoforms, CYP3A4 has been recognized as one of the most important CYP enzymes, owing to its high abundance in metabolic organs and the extremely broad substrate spectra.3 As a key contributor to xenobiotic metabolism, CYP3A4 catalyzes the oxidative metabolism of numerous toxicological or pharmacological agents including carcinogens, environmental pollutants, food chemicals, phytochemicals and marketed drugs.3–6 Dysfunction or strong inhibition of CYP3A4 may affect the metabolic clearance of CYP3A4 substrate drugs, which may cause pharmacokinetic drug-drug interactions (DDIs) and result in adverse drug reactions (ADRs).7 To avoid harmful DDIs, the regulatory agencies including the United States Food and Drug Administration (USFDA) and the European Medicines Agency (EMA) have recommended to assess the DDI potentials of drug candidates against the pivotal hepatic CYPs (such as CYP3A4) before drug approval, which emphasized the importance of novel and efficient tools for predicting CYP3A4 inhibition potentials of the drug candidates.

Over the past few decades, a panel of experimental approaches have been developed to screen CYP3A4 inhibitors, which primarily rely on quantitative high-throughput screening methods including bioluminescence or fluorescence-based assays, as well as mass spectrometry-based assays.8–10 However, these experimental methods still require a significant investment of skilled operators, high-cost devices and redundant sample preparation procedures. To address these challenges, computational methods have emerged as valuable tools for high-throughput assessing biological activities of compounds in the early stages of drug discovery.11 Virtual screening, in particular, has the potential to save a significant amount of time and resources in the lead selection and structure optimization. Nevertheless, canonical structure-based docking methods have encountered difficulties due to the inherent flexibility of CYP structures.12–14 On the other hand, data-driven approaches have demonstrated superior prediction accuracy for CYP inhibition activity.15–17 In recent years, the accumulation of data,18,19 advancements in computing hardware, and the rapid development of deep learning (DL) architectures have strongly facilitated the development of various computational tools for predicting the inhibition potentials of CYP enzymes, particularly 1A2, 2C9, 2C19, 2D6 and 3A4.15,20–22 A recent review has summarized the CYP metabolism prediction tools developed since 2019, including well-known ADMET tools such as HelixADMET,23 ADMETLab 2.0,24 Interpretable-ADMET,25 FP-ADMET,26 and admetSAR 2.0,27 as well as specialized prediction tools for CYPs, such as DeepCYP,20 CYPlebrity,15 SuperCYPsPred,28 and iCYP-MFE.29 However, all current tools are deterministic classification models designed to predict whether a compound acts as an inhibitor or not for the target CYP enzyme(s), without assessing the reliability of these predictions

The key to the success of machine learning (ML) models is the availability of comprehensive and high-quality datasets. In the field of molecular property prediction, bioactive data are inherently noisy due to various reasons such as diverse experimental methods, experimenter manipulation, and instrumental resolution.30 As a result, data quality from different sources can vary significantly, leading to a potential impact on the coherence of information and ultimately affecting the performance of the model.30,31 In addition, limited availability of training data often leads to poor generalization capabilities of models. Inconsistent distributions between the training and test data can result in models that generate unreliable outputs, often underestimating the actual errors that users may encounter when applying the model to predict new compounds in real-world scenarios.30,31 Unfortunately, traditional DL models do not offer confidence estimates for their outputs. In regression tasks, the output is a singular deterministic value, while in classification tasks, the output is a probability distribution with inadequate calibration of confidence. The limitation of reliability can have a detrimental impact on the decision-making process when utilizing prediction tools for drug design.30,32

Uncertainty quantification (UQ) methods are promising approaches that enable estimating the reliability of prediction results. Generally, uncertainty can be classified into two categories: aleatoric uncertainty (data-related, irreducible) and epistemic uncertainty (model-related, reducible). Aleatoric uncertainty reflects the inherent stochastic nature of the data. In the field of drug discovery, data noise arises from both systematic and random errors in experimental measurements.33 Aleatoric uncertainty cannot be reduced by dataset expansion or improving model performance. Otherwise, it can be reduced by utilizing consistent experimental sources, replicating experimental results, and removing erroneous samples. Epistemic uncertainty refers to errors arising from lacking knowledge in certain regions of the sample space.31–33 It is determined by the distribution of model parameters and reflects the confidence level of the model. Typically, higher epistemic uncertainty is assigned to predictions in regions with sparse data points, while lower epistemic uncertainty is assigned in regions with observed data points.30 To improve epistemic uncertainty, expanding samples in low-density data areas and enhancing the model's predictive capabilities can be beneficial. Over the past few decades, significant progress has been made in the development of various strategies for UQ, such as Bayesian approximation methods (e.g., Monte Carlo dropout,34 Markov Chain Monte Carlo (MCMC),35 Variational inference (VI),36,37 Bayes By Backprop (BBB),38 Gaussian Approximation39), ensemble methods (e.g., Deep Ensembles,40 Gradient Boosting,41 Bagging,42 Bayesian Model Averaging (BMA)43), and conformal prediction methods (e.g., Inductive Conformal Prediction (ICP),44 Transductive Conformal Prediction (TCP)45). Amini et al.46,47 proposed an efficient Evidential Deep Learning method without sampling, making it an attractive alternative. These methods have been explored in various applications, including molecular property prediction, target prediction, and chemical reaction prediction, showing promising potential. However, the practical effectiveness of UQ methods still requires further evaluation and validation.

This study aims to establish a novel uncertainty-guided DL method for screening CYP3A4 inhibitors. In particular, we address the following issues to make the approach in practice. We first collected and integrated a comprehensive dataset of CYP3A4 inhibitors involving 27,045 compounds with classification labels or regression values. Considering that accurate prediction of inhibitory intensity can provide better guidance for experiments in practical applications, we intend to build a regression model that incorporates the classification outcomes to forecast IC50 values. We demonstrate that the classification model based on a large dataset significantly improves the performance of the regression model trained on a relatively smaller dataset. Furthermore, the evidential-based UQ method was introduced to assess the confidence of prediction values. The results show that the uncertainty method could effectively calibrate the error distribution and accurately characterize the correlation between uncertainty and prediction errors. Consequently, predictions with higher confidence exhibit better performance. To showcase the practicality of our new approach, we conducted an uncertainty-guided virtual screening process of an in-house compound set and validated the screening results through *in vitro* inhibition assays. As a result, within the top 20 compounds ranked with uncertainty, a total of 15 compounds were identified as new CYP3A4 inhibitors. Notably, compound 4'-methoxy-2'-hydroxy quinoxaline-chalcone, 2’,4’-dihydroxychalcone, raloxifene exhibited potent anti-CYP3A4 activities, with IC50 values of 0.428 μM, 0.497 μM, and 0.918 μM, respectively.

# 2. MATERIALS AND METHODS

2.1. Data collection and curation

Chemical data was obtained from the PubChem Bioassay Database and the ChEMBL Database for model development and evaluation. Previous studies primarily utilized a dataset composed of 17,143 samples collected from PubChem Assay ID (AID) 1851.15,20,28,29 To enhance dataset diversity, we collected additional 11 datasets, mainly containing AID 884, AID 1645841, AID 1671201, and AID 686941. Each sample in the dataset was labeled as "active", "inactive", and "inconclusive". To ensure data quality and consistency, "inconclusive" samples were excluded, and only "active" (inhibitors) and "inactive" (non-inhibitors) compounds were retained for further processing. The cutoff for labeling "active" and "inactive" compounds was set at pIC50=5. For data collection from the ChEMBL database, we obtained data with exact pIC50 labels. Data labeled as "Outside typical range" were removed, and the remaining data with pIC50 values were kept for subsequent processing.

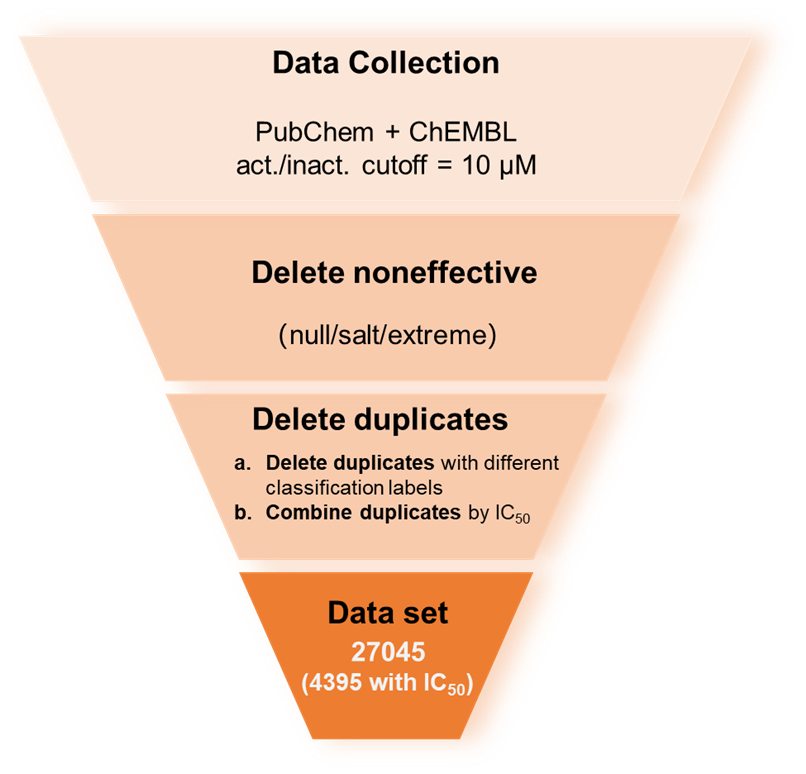


Figure 1. Flowchart of CYP3A4 inhibitors data processing.

All SMILES in the dataset were standardized to ensure consistency and facilitate analysis. This process involved removing fragments, metal ions, and stereochemical information, as well as converting to the Canonical SMILES format. We categorized sample conflicts into two types: (1) compounds with the same structure but different activity labels, and (2) compounds with the same structure and the same activity label. For data with classification labels, we removed all conflicted samples in case (1) and deduplicated samples in case (2) to obtain a unique representative. For compounds with specific IC50 values, we deduplicated data in case (2) to own a unique label. For case (1), the same compounds with different IC50 values often occur in real experiments due to the utilization of consistent positive controls, but different experimental methods may produce different IC50 values. We performed greedy algorithm to merge activity values across different experimental groups while maintaining the sorting within each group unchanged. The detailed processing procedure is illustrated in Figure S1.

After processing, a total of 27,045 compounds with classification labels were collected, of which 4,395 had explicit IC50 values (Figure 1).

## 2.2 Model Construction and Optimization

### 2.2.1 Molecular Representation

Molecular fingerprints are vectors that encode molecular structural characteristics, extensively used in drug design, virtual screening, and similarity comparisons.26,48,49 Different types of molecular fingerprints are suitable for specific tasks, and exhibit varying performance under different requirements.50 The selection of appropriate fingerprints for ML models significantly impacts their performance and generalizability. Here, we systematically evaluated the model performance of 15 popular fingerprints, including linear connectivity fingerprints (RDKit, AP2D, APC, EStateFP), circular fingerprints (ECFP6, FCFP4, Morgan), and substructure fingerprints (MACCS, PubChem, SubFP, SubFPC) (Table S1). All fingerprints were calculated by PaDEL package (Version 3) and RDKit (Version 2022.9.5).

Moreover, recognizing that standalone molecular fingerprints inherently capture certain information of chemical molecules, we explored the performance of coupling fingerprints obtained by combining two fingerprints in three different ways: (1) Feature selection by Group-Lasso (GL) algorithm51: The extension of the Least Absolute Shrinkage and Selection Operator (LASSO) simultaneously performs individual feature selection and group-level sparsity, effectively identifying the most informative features and relevant groups to enhance ML models. (2) Pearson correlation coefficient (PCC) evaluation: PCC is calculated for each fingerprint pair, and the least related groups are combined to create a complementary and non-redundant coupling fingerprint. (3) Concatenation of the top-performing single fingerprints (TOP): The two fingerprints with the best individual performance are combined, aiming to leverage their strengths for potentially improved predictions.

### 2.2.2 Model Architecture

The model framework comprises three key components: the Classification module, the Regression module, and the Evidential layer (Figure 2). In the Classification module, molecular fingerprints served as input and underwent feature extraction and classification through two convolutional blocks and one fully-connected (FC) block. Then, the probability predicted by the classification model was determined by the output of the last FC layer and activated using the Sigmoid function. Similarly, in the Regression module, molecular fingerprints were utilized as inputs, and feature extraction and regression were carried out through two convolutional blocks and one FC block. The predicted labels from the Classification module were concatenated with the inputs of the Regression module and fed into its second FC layer. Other concatenation methods were also performed and evaluated (Figure S2 and Table S7). The final layer of the regression model, known as the Evidential layer, was derived from the Evidential Learning method46 and encompassed four parameters denoted as . The evidence loss can be expressed as:

Here, represents a set of weights. The log-likelihood term is utilized to measure model fit, while controls the penalty strength of overconfident predictions by the regularization term .

The prediction value, aleatoric, and epistemic uncertainty could be calculated using parameters as follow:

, ,

### 2.2.3 Model Optimization

For model training and evaluation, the regression dataset was divided into training, validation, and testing sets using random sampling to maintain a consistent distribution of data (Table 1). We emphasized that the regression test data with 880 compounds were kept separate and not included in either the classification or regression training/validation sets, serving as an independent test set. It allowed for an unbiased evaluation of the regression model, classification-combined regression model, and regression model with uncertainty estimation. Hyperparameters were optimized using grid search, and the final settings were bolded in Table 2.

Table 1. Data splitting for model development and evaluation.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Dataset | Data class | Training | Validation | Test | Total |
| Classification | noninhibitor | 9196 | 3066 | 3065 | 15327 |
| inhibitor | 7031 | 2344 | 2343 | 11718 |
| total | 16227 | 5410 | 5408 | 27045 |
| Regression | — | 2812 | 703 | 880 | 4395 |

Table 2. Hyperparameters settings.

|  |  |
| --- | --- |
| Hyperparameters | Values*\** |
| Regularization coefficient λ | 0.005, 0.01, 0.05, 0.1, **0.2**, 0.4, 0.6, 0.8,1 |
| Batch size | 16, 32, **64**, 128, 256 |
| Learning rate | 10-1**, 10-2**, 10-3, 10-4 |
| Dropout | **0.2**, 0.4, 0.6 |

*\** Best hyperparameters are marked as bold.

2.3 Model Evaluation

For classification models, the area under the receiver operating characteristic (ROC) curve (AUC-ROC), Accuracy (ACC), balanced accuracy (BACC), precision (PR), recall (RE), specificity (SP), F1 score (F1), and Matthews’s correlation coefficient (MCC) were used to access model ability. For the regression model, Root mean square error (RMSE), Pearson correlation coefficient (R) and R square (R2) were calculated to evaluate model performance. All metrics were calculated by Scikit-learn (Version 1.0.2).

2.4 CYP3A4 inhibition activity assay

The procedure and details of CYP3A4 inhibition assays were followed as described in Zhang et al.'s work.52 In short, a highly specific CYP3A4 substrate, N-ethyl-1,8-naphthalimide,53 was employed to evaluate the residual activities of CYP3A4 in the presence of test compounds. Ketoconazole served as the positive control, while acetonitrile was used as the blank control. All measurements were performed in triplicate. IC50 values were calculated using nonlinear regression analysis with Prism version 8.3.0 (GraphPad Software, CA, USA).

2.5 Molecular Docking

AutoDock Vina (Version 1.1.2) was used to perform docking between ligands and CYP3A4. The ligand was initially provided in SMILES format and underwent 3D molecular conformation optimization using the Python RDKit toolkit. Subsequently, it was converted to PDBQT format using Open Babel (Version 3.1.1). The protein structure of CYP3A4 was retrieved from RCSB Protein Data Bank (PDB) (PDB ID: 3NXU) and underwent several structure processing steps, including the removal of water molecules, the addition of hydrogen atoms, and application of Kollman charges. The grid box covering the entire active site pocket was chosen. Only the lowest docking score (the best one) was used for subsequent analysis. The docking results of the ligand and CYP3A4 were analyzed and visualized using Discovery Studio Visualizer (BIOVIA Discovery Studio 2019, Dassault Systèmes, SanDiego, USA).

# 3. RESULTS AND DISCUSSIONS

Accurate prediction of CYP3A4 inhibitory activity is crucial in experiments. Previously reported models for CYP inhibitors prediction have primarily focused on binary classification,15,20,29 distinguishing between compounds with and without inhibitory activity using a threshold such as IC50 at 10 μM. In the actual drug discovery process, it is essential to predict the inhibitory intensity and confidence level to provide better guidance for drug design. To address this issue, we collected a comprehensive dataset of CYP3A4 inhibitors and developed an uncertainty-guided regression model for predicting CYP3A4 inhibitory activity (Figure 2). The model construction contains two main steps. In the first step, a robust classification model was developed by leveraging all binary classification data. Then, the classification outcomes were integrated with the input features of the regression model. In the second step, an evidential uncertainty method was employed to quantify the prediction confidence by estimating the aleatoric and epistemic uncertainty. Finally, we applied the model to conduct virtual screening based on the different ranking strategies, and validated the predicted inhibition values.

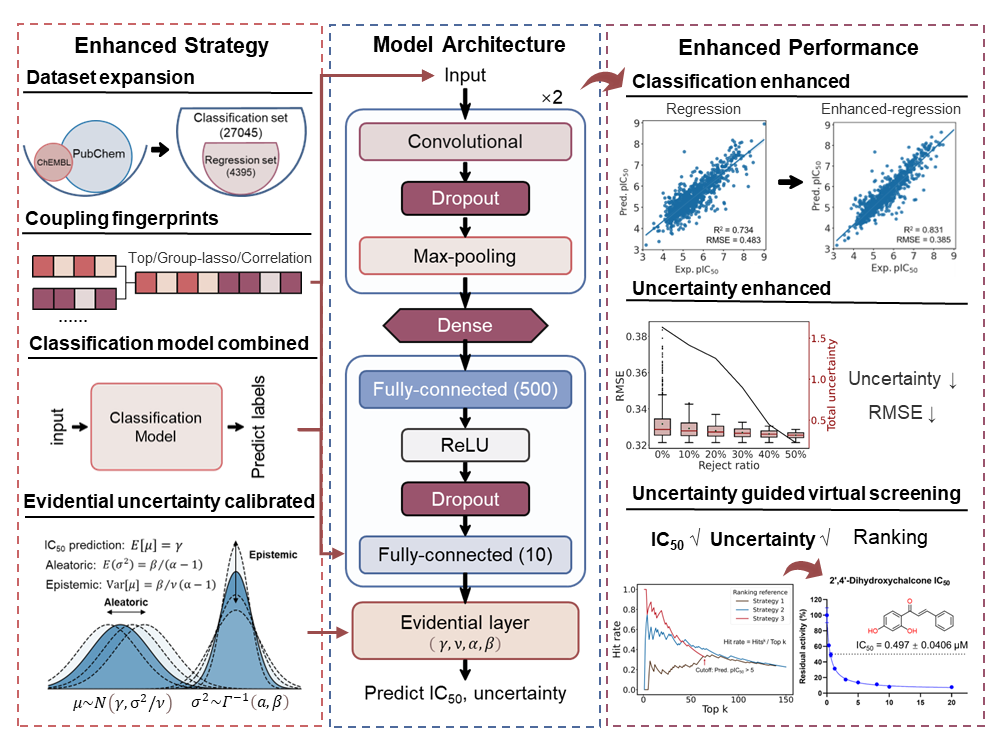


Figure 2. Enhanced strategy, model architecture and performance evaluation.

* 1. Physicochemical properties and structure analysis

In this work, we have compiled the most comprehensive and extensive dataset for CYP3A4 inhibitors to date. The dataset comprises 27,045 compounds with classification labels, in which 4,395 compounds have the IC50 values against human CYP3A4. To comprehensively analyze this dataset, the data were divided into three groups based on the known labels, namely strong inhibitor group (IC50 < 1 μM), active inhibitor group (IC50 < 10 μM) and inactive group (IC50 > 10 μM).

As shown in Figure 3A and Table S2, all six physicochemical properties of the compounds within three different groups exhibited significantly different (Mann–Whitney U test, p<0.001). The strong inhibitors displayed larger molecular weights compared to the inactive and active compounds, which could be attributed to the large active pocket of CYP3A4.12,13 Moreover, the strong inhibitor and active inhibitor groups generally had more hydrogen bond acceptors than the inactive compounds, agreeing well with the fact that potent CYP3A4 inhibitors prefer to form hydrogen bonds with amino acids around the CYP3A4 catalytic pocket.12 Notably, the strong inhibitors exhibited relatively larger MolLogP values, which is consistent with the ligand preference of CYP3A4 (which prefers to bind the ligand bearing polyaromatics that always showed high hydrophobicity).

We further visualized the chemical structure space of three inhibitor groups using t-SNE plots (Figure 3B). t-SNE is an algorithm used to reduce high-dimensional data into a two-dimensional representation, highlighting patterns and similarities among data points. The plots showed a substantial overlap between the chemical structure spaces of the active and inactive groups, suggesting no significant differences in chemical structure among these groups. Moreover, the strong inhibitors exhibited a relatively uniform distribution within the active group’s space, indicating the chemical structural diversity among the potent CYP3A4 inhibitors.

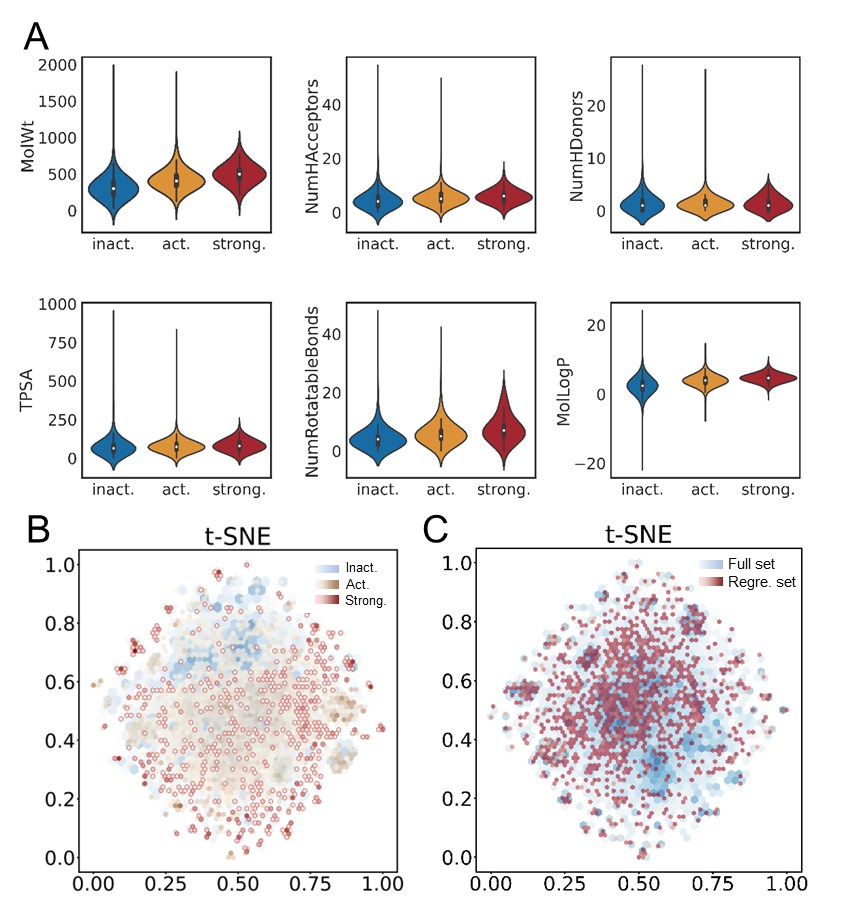
****

Figure 3. Comparison of physicochemical features in strong, active and inactive compounds. (A) Box plots of the six physicochemical properties for the three groups. (B) Density plots illustrate the distribution of chemical structures for the three groups (inactive, active, and strong inhibitors) with blue, yellow, and red colors, respectively. Darker colors indicate higher compound density at specific points. (C) Density plots shows the distribution of chemical structures for the complete dataset (blue) and the regression dataset (red). Darker colors indicate higher compound density at specific points.

3.2 Classification-enhanced regression model performance

Accurate prediction of inhibition activities is crucial for guiding drug design and development processes. Here, we constructed a regression DL model based on 4,395 data with IC50 values (see Figure S2 for the regression model architecture). To effectively utilize the diverse molecular fingerprints, including substructure-based fingerprints (e.g., MACCS, PubChem fingerprints), linear topology-based fingerprints (e.g., RDKit fingerprints), and circular topology-based fingerprints (e.g., Morgan, FCFP4 fingerprints), we trained the regression model using 15 standalone molecular fingerprints and 3 coupling fingerprints. The coupling fingerprints were generated using three splicing strategies, as detailed in the "Materials and Methods" section. Our results indicated that the TOP coupling fingerprint yielded the best performance with RMSE of 0.483 and R2 of 0.734 (Table S3 and Table 3). These findings highlight the effectiveness of coupling fingerprints by leveraging diverse molecular fingerprints.

Given a significantly larger set comprising 27,045 classification data and the greater chemical diversity observed in the t-SNE plot (Figure 3C), we proposed to improve the regression performance by incorporating the classification results. In parallel, we developed a classification predictive model using the coupling fingerprint and convolutional neural network (CNN) architecture (see Figure S2 for classification model architecture and Table S4, Table S5, and Figure S3 for classification model results). To assess whether the classification model trained on the larger dataset exhibits better performance and generalization ability, we evaluated both the regression model and the classification model on the regression independent test set of 880 samples. Notably, the classification model exhibited a remarkable improvement in accuracy (ACC = 0.794) compared to the regression model (ACC = 0.547, wherein the regression results were transformed into binary classification using a threshold of IC50 10 μM) (Table S6).

Encouraged by these promising results, we proceeded to integrate the classification outcomes into the regression model using various strategies (see Figure S2 for three types of model concatenating strategies and Table S7 for the corresponding results). The best performing strategy, as shown in Figure 2, demonstrated that the classification-combined regression model outperformed the regression CNN model, achieving RMSE of 0.385 and R2 value of 0.831 (Table 3). Moreover, the scatter plot shown in Figure 4 depicted a more density distribution, indicating that the classification model partially corrected the regression predictions, thereby reducing regression errors. These findings underscore the potential of leveraging a larger classification dataset to enhance regression models and provide valuable insights for biological activity modeling.

Table 3. Performance of regression models and the classification-combined regression model.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Splicing Mode*\** | | RMSE | R | R2 | Epistemic | Aleatoric | Total |
| **TOP** | | **0.483** | **0.858** | **0.734** | **0.337** | **0.279** | **0.616** |
| GL | | 0.711 | 0.655 | 0.428 | 0.526 | 0.486 | 1.012 |
| PCC | | 0.712 | 0.657 | 0.427 | 0.677 | 0.569 | 1.246 |
| **CL + TOP** | **0.385** | **0.916** | **0.831** | **0.234** | **0.219** | **0.453** |

*\**TOP refers to the top-performing two single fingerprints splicing mode. GL refers to the most informative fingerprint by Group Lasso splicing mode. PCC refers to the least Pearson correlation coefficient features splicing mode. CL + TOP refers to the Classification-combined top-performing regression model. Best results are marked as bold.

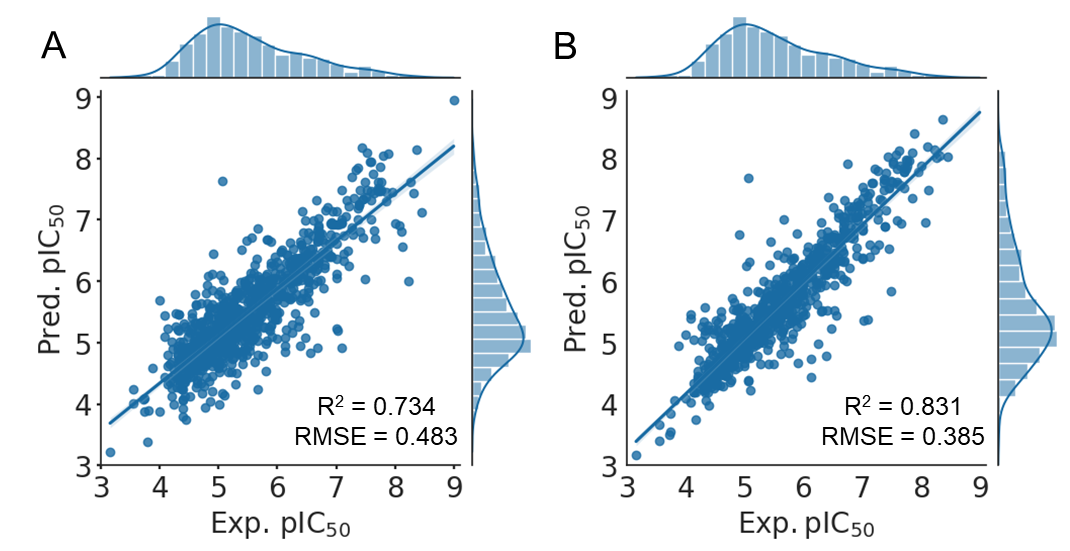
****

Figure 4. Model performance of (A) the top-performing regression model and (B) the Classification-combined top-performing regression model on the test set.

3.3 Uncertainty-enhanced model performance

In addition to improving predictive accuracy, obtaining accurate confidence estimates for model predictions is crucial for guiding experimental decisions. To this end, we incorporated the evidential uncertainty method as a post-processing layer to the predictive neural network. The evidential uncertainty method involves introducing evidential priors on the original Gaussian likelihood function and training the neural network to infer the hyperparameters of the evidential distribution so that both aleatoric and epistemic uncertainty can be learned. Previous works have demonstrated its capability to capture the domain of out-of-distribution (OOD) samples and provide well-calibrated uncertainty estimates, resulting in precise estimation of the error distribution's variance and confidence interval.46 This approach has proven effective in various fields, including image recognition and molecular property prediction.46,47,54 Moreover, the evidential method is user-friendly, as it does not requiresampling and can be implemented without significant architecture changes.32,46

Here, we evaluated the calibration and ranking capabilities of the evidential model. The evidence loss function includes two terms, one for maximizing evidence and the other for regularization, which is scaled by a regularization coefficient, λ. By varying λ, the method's sensitivity to OOD samples can be adjusted.46,47,55 A perfect uncertainty estimator would result in an observed distribution that exactly matches the ideal calibration. Generally, smaller values of λ yield over-confident estimates, while higher λ values may cause under-confidence. As shown in Figure 5 A, a value of λ equal to 0.2 provided well-calibrated uncertainty quantification and was selected for subsequent analysis.

The distribution of total uncertainty, epistemic uncertainty, and aleatoric uncertainty of test data was analyzed. Their values ranged from 0.22 to 1.63, with aleatoric uncertainty exhibiting a more concentrated distribution (Figure 5B). To assess the ranking ability of the evidential uncertainty, where higher uncertainty values should be assigned to larger prediction errors, we conducted independent sample t-tests on the absolute error values of the high uncertainty value group (top 25% of uncertainty in test data) and the low uncertainty group (bottom 25% of uncertainty in test data). Figure 5C showed significant differences between the two groups for all three uncertainty calculation methods, with p-values less than 0.001. Subsequently, we sorted the data based on the total uncertainty and removed data that exceeded a certain percentage. As depicted in Figure 5D, the RMSE rapidly decreased with increasing confidence, indicated that high-confidence predictions can significantly enhance the model's performance. Similarly, Figure 5E showed that as the uncertainty values increased, the proportion of data with larger absolute errors also increased, further confirming the strong correlation between evidential uncertainty and prediction errors.

To visualize the uncertainty distribution in chemical space, we projected the test compounds into t-SNE plot (Figure 5F). The size of each circle corresponds to the magnitude of the prediction error, while the color represents the uncertainty value. We observed that most points with larger errors exhibited higher uncertainty values. The distribution of points with different uncertainty values and prediction errors in chemical space were widely dispersed, with only a few clusters, suggested that the uncertainty estimates were not biased towards specific chemical structures.

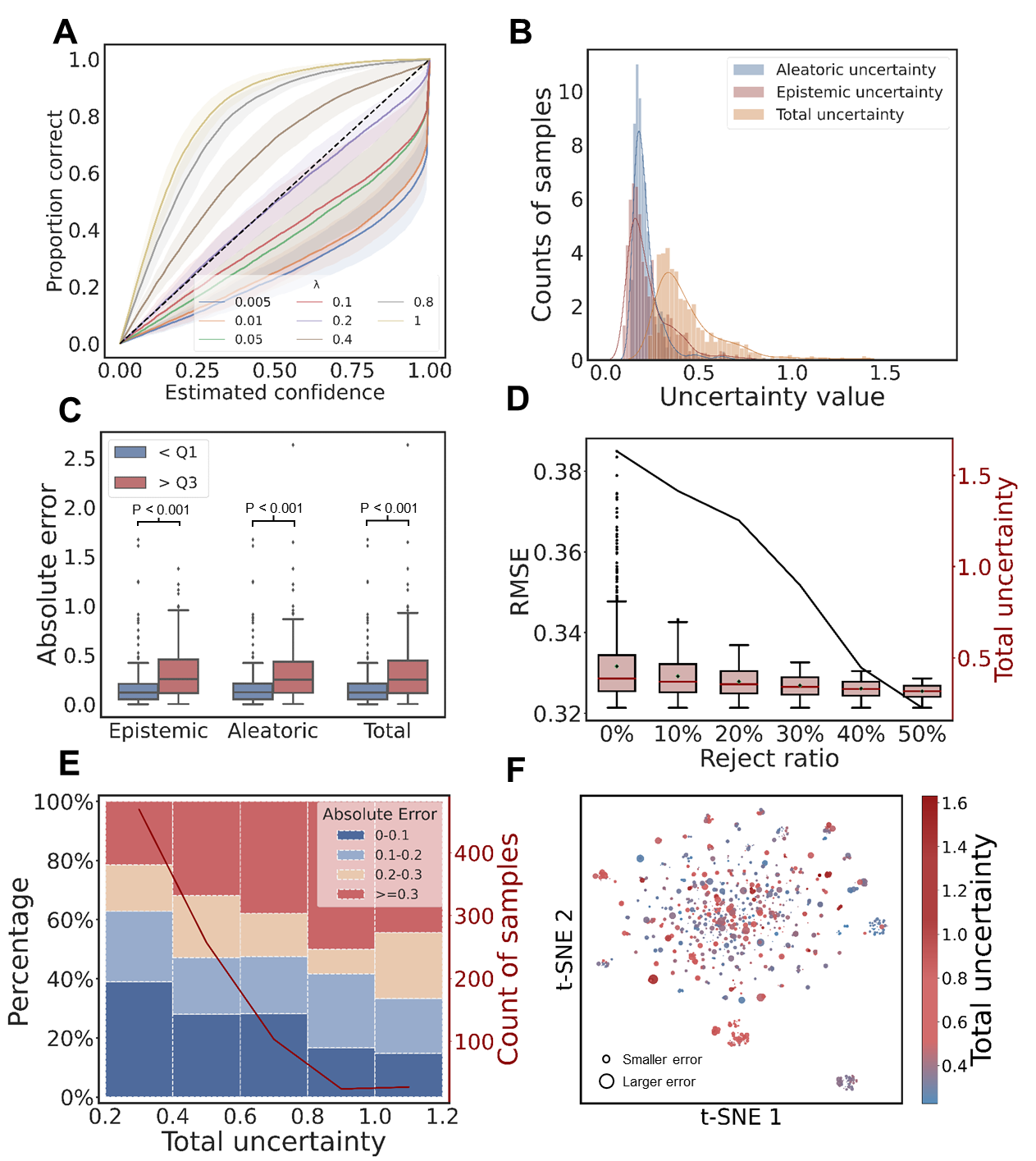


Figure 5. Uncertainty analysis using evidential uncertainty-based regression model. (A) Estimated confidence against the observed proportion correct evaluated on test set. The dashed line represents a perfect calibration. Mean ± 95% c.i., n =10 independent trials. (B) Histograms of the distribution of the aleatoric, epistemic, and total uncertainty. (C) Student’s t-tests on absolute error distributions for aleatoric, epistemic, and total uncertainty in two groups of the test set: Q1 (top 25% uncertainty values) and Q3 (bottom 25% uncertainty values). All p-values < 0.001. (D) The test data was sorted and sequentially rejected based on decreasing total uncertainty percentiles, and RMSE (black line) for the remaining data was calculated. Boxplots represent the distribution of uncertainty values for data at different uncertainty percentile cutoffs. (E) Percentage of data with different absolute errors at various uncertainty intervals evaluated on the test set. Red, yellow, light blue, and blue represent the cutoffs of absolute errors 0-0.1, 0.1-0.2, 0.2-0.3, and ≥0.3, respectively. The red line represents the number of samples in each of the five intervals. (F) Visualization of total uncertainty in the chemical structural space on the test set. Circle size indicates the magnitude of the error, while color represents the uncertainty value.

3.4 Uncertainty-guided virtual screening and experimental validation

Based on these promising results, we conducted uncertainty-guided virtual screening of CYP3A4 inhibitors using 150 in-house compounds that were not used in model construction. We predicted the pIC50, as well as aleatoric and epistemic uncertainty values for each compound. To comprehensively evaluate the 150 predictions, we performed single-point screening for all compounds, and the active compounds were further subjected to accurate IC50 measurements (Table S8).

The hit rate of virtual screening is a crucial performance metric frequently used to evaluate the effectiveness of computational methods in identifying active compounds from a large chemical database. Here, we compared the hit rate based on three ranking strategies: Strategy 1 (ranking by predictive pIC50 values), Strategy 2 (ranking by total uncertainty values), and Strategy 3 (ranking data with predictive pIC50 > 5 based on total uncertainty values). Interestingly, uncertainty-informed ranking (Strategies 2 and 3) led to a significant increase in the hit ratio compared to only considering pIC50 values (Strategy 1). Notably, Strategy 3 showed the most remarkable enrichment, with 8 hits found in the top 10 rankings and 15 hits found in the top 20 rankings (Figure 6, Table 4, and Table S9). At the same time, it was observed that data with lower uncertainty values had lower prediction errors (Table 4). This result highlights the importance of ranking with uncertainty, as it substantially enriches the candidate set for experimental hits, while the performance is greatly reduced when only considering predictive value without uncertainty.

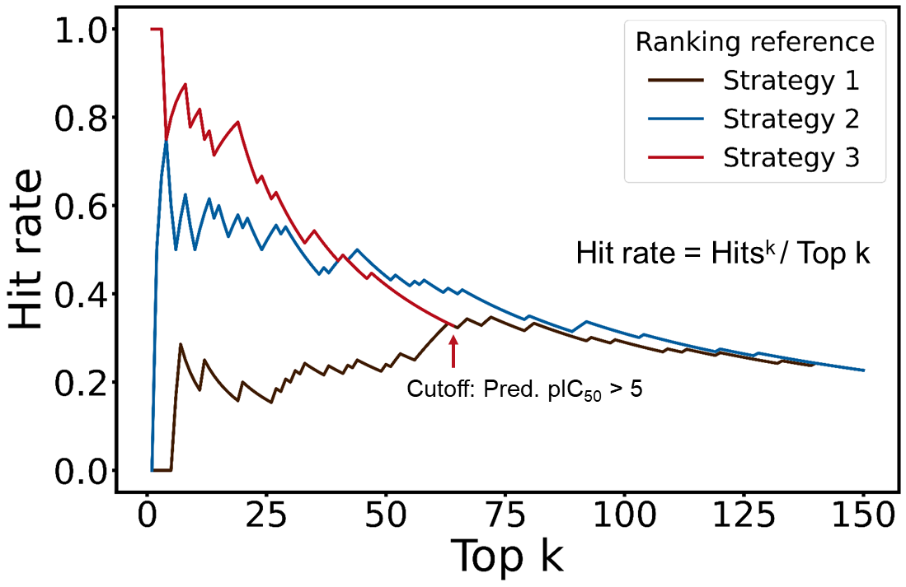


Figure 6. The hit rate of virtual screening among 150 in-house compounds. Strategy 1 indicates ranking by predictive pIC50 values, Strategy 2 indicates ranking by total uncertainty values, and Strategy 3 indicates ranking data with predictive pIC50 > 5 by total uncertainty values.

Table 4. Hit rates and RMSEs of the top-ranking predictions by Strategy 3a.

|  |  |  |  |
| --- | --- | --- | --- |
| Top k | Hit number | Hit rate | RMSEb |
| **top 10** | **8** | **0.800** | **0.538** |
| top 20 | 15 | 0.750 | 0.667 |
| top 30 | 17 | 0.567 | 0.733 |

aBest results are marked as bold.

b RMSE is calculated on data that has pIC50 values.

Based on Strategy 3, we identified 15 new CYP3A4 inhibitors from the top 20 compounds. Notably, 4'-methoxy-2'-hydroxy quinoxaline-chalcone, 2’,4’-dihydroxychalcone, raloxifene exhibited activity less than 1 μM, with IC50 values of 0.428 μM, 0.497 μM, and 0.918 μM, respectively.

Molecular docking was utilized to reveal the binding modes of these compounds toward CYP3A4. The docking scores for 4'-methoxy-2'-hydroxy quinoxaline-chalcone, 2’,4’-dihydroxychalcone, and raloxifene were -8.369 kcal/mol, -7.715 kcal/mol, and -9.608 kcal/mol, respectively, indicating strong interactions with CYP3A4. We observed that all three compounds contain benzene or heterocyclic rings, which can form π-π interactions with the heme group in active cavity of CYP3A4, suggesting that the aromatic rings play a key role in CYP3A4 inhibition. Additionally, Glu374 was found to form hydrogen bonds with the hydroxyl group of 2’,4’-dihydroxychalcone and form weak hydrogen bonds with the other two compounds, indicating that hydrogen bonding also contributes to the potent inhibition (Figure S4). These findings align with the preferred physicochemical properties of potent inhibitors.

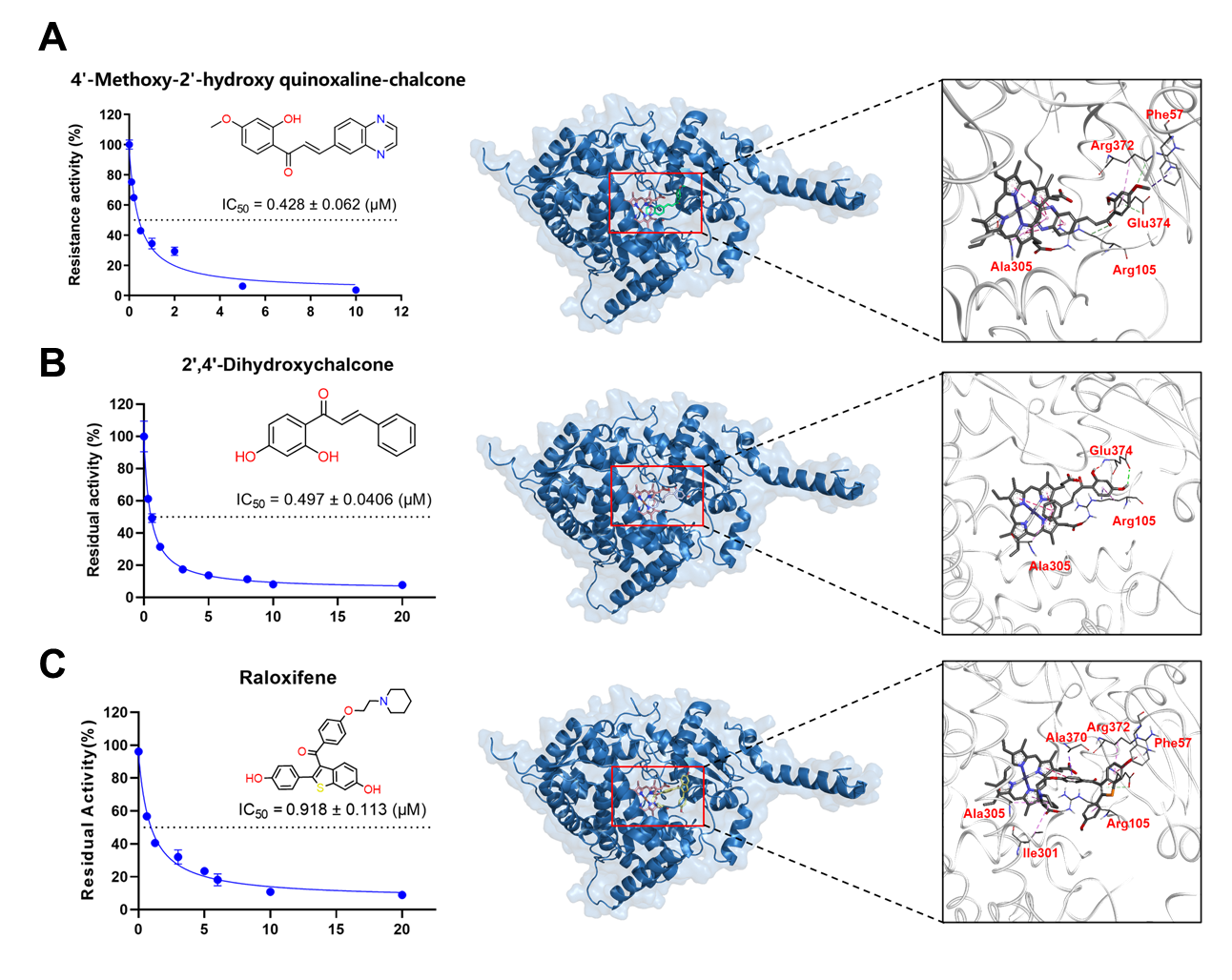
****

Figure 7. Experimental IC50 and interaction analysis of three potent inhibitors: (A) 4'-methoxy-2'-hydroxy quinoxaline-chalcone, (B) 2’,4’-dihydroxychalcone, and (C) raloxifene. (Left) IC50 curves. (Right) The stereo and detailed interactions.

# 4. CONCLUSIONS

In this paper, we present a robust, reliable, and efficient regression model for predicting CYP3A4 inhibition intensity. The model's performance is enhanced by the utilization of a larger chemical space, the combination of diverse molecular fingerprints, the integration of classification model results, and the incorporation of uncertainty estimates. We emphasize the importance of evidential uncertainty estimation, which helps calibrate prediction results and boosts confidence in the accuracy of predictions. To validate our model's efficacy, we applied it to screen an in-house compound set and confirmed 8 hits out of the top 10 compounds with uncertainty ranking, demonstrating a significantly higher hit ratio than using prediction values alone. These results highlight the effective guidance of evidential uncertainties in accelerating virtual screening. In future computational prediction-guided experiments, we strongly recommend prioritizing compounds with higher certainty for experimental validation.

# DATA AND SOFTWARE AVAILABILITY

All data involved in this study and the source code are available at <https://github.com/wangwrx/An-uncertainty-guided-deep-learning-method-facilitates-rapid-screening-of-CYP3A4-inhibitors>.

# ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/xx.xxx/acs.jcim.xxxxx>.

# AUTHOR INFORMATION

**Corresponding Authors**

Guangbo Ge − Shanghai Frontiers Science Center of TCM Chemical Biology, Institute of Interdisciplinary Integrative Medicine Research, Shanghai University of Traditional Chinese Medicine, Shanghai 201203, China. E-mail: [geguangbo@shutcm.edu.cn](mailto:geguangbo@shutcm.edu.cn)

Xiaoqing Guan − Shanghai Frontiers Science Center of TCM Chemical Biology, Institute of Interdisciplinary Integrative Medicine Research, Shanghai University of Traditional Chinese Medicine, Shanghai 201203, China. E-mail: [guanxq@shutcm.edu.cn](mailto:guanxq@shutcm.edu.cn)

Qingping Zhou − School of Mathematics and Statistics, Central South University, Changsha 410083, Hunan, China. E-mail: [qpzhou@csu.edu.cn](mailto:qpzhou@csu.edu.cn)

**Authors**

Ruixuan Wang − Shanghai Frontiers Science Center of TCM Chemical Biology, Institute of Interdisciplinary Integrative Medicine Research, Shanghai University of Traditional Chinese Medicine, Shanghai 201203, China

Zhikang Liu − School of Mathematics and Statistics, Central South University, Changsha 410083, Hunan, China

Jiahao Gong − Shanghai Frontiers Science Center of TCM Chemical Biology, Institute of Interdisciplinary Integrative Medicine Research, Shanghai University of Traditional Chinese Medicine, Shanghai 201203, China

Complete contact information is available at: https://pubs.acs.org/doi/xx.xxx

**Notes**

The authors declare no competing financial interest.

**ACKNOWLEDGMENTS**

This work was financially supported by the NSF of China (82204606, 81922070, 81973286, 82003847, and 12101614), Shanghai Municipal Health Commission's TCM Research Project (2022CX005), Innovation Team and Talents Cultivation Program of National Administration of Traditional Chinese Medicine (ZYYCXTDD-202004), the Natural Science Foundation of Hunan Province. China (Grant No.2021JJ40715).

**REFERENCES**

(1) Zhao, M.; Ma, J.; Li, M.; Zhang, Y.; Jiang, B.; Zhao, X.; Huai, C.; Shen, L.; Zhang, N.; He, L.; Qin, S. Cytochrome P450 Enzymes and Drug Metabolism in Humans. *Int J Mol Sci* **2021**, *22* (23), 12808.

(2) Evans, W. E.; Relling, M. V. Pharmacogenomics: Translating Functional Genomics into Rational Therapeutics. *Science* **1999**, *286* (5439), 487–491.

(3) Guengerich, F. P. Cytochrome P450s and Other Enzymes in Drug Metabolism and Toxicity. *AAPS J* **2006**, *8* (1), E101–E111.

(4) Guengerich, F. P.; Shimada, T. Oxidation of Toxic and Carcinogenic Chemicals by Human Cytochrome P-450 Enzymes. *Chem. Res. Toxicol.* **1991**, *4* (4), 391–407.

(5) Basheer, L.; Kerem, Z. Interactions between CYP3A4 and Dietary Polyphenols. *Oxidative Medicine and Cellular Longevity* **2015**, *2015*, e854015.

(6) Choi, J. S.; Piao, Y. J.; Kang, K. W. Effects of Quercetin on the Bioavailability of Doxorubicin in Rats: Role of CYP3A4 and P-gp Inhibition by Quercetin. *Arch. Pharm. Res.* **2011**, *34* (4), 607–613.

(7) Deodhar, M.; Al Rihani, S. B.; Arwood, M. J.; Darakjian, L.; Dow, P.; Turgeon, J.; Michaud, V. Mechanisms of CYP450 Inhibition: Understanding Drug-Drug Interactions Due to Mechanism-Based Inhibition in Clinical Practice. *Pharmaceutics* **2020**, *12* (9), 846.

(8) Ge, G. B.; Ning, J.; Hu, L. H.; Dai, Z. R.; Hou, J.; Cao, Y. F.; Yu, Z. W.; Ai, C. Z.; Gu, J. K.; Ma, X. C.; Yang, L. A Highly Selective Probe for Human Cytochrome P450 3A4: Isoform Selectivity, Kinetic Characterization and Its Applications. *Chem Commun (Camb)* **2013**, *49* (84), 9779–9781.

(9) Qi, X. Y.; Liang, S. C.; Ge, G. B.; Liu, Y.; Dong, P. P.; Zhang, J. W.; Wang, A. X.; Hou, J.; Zhu, L. L.; Yang, L.; Tu, C. X. Inhibitory Effects of Sanguinarine on Human Liver Cytochrome P450 Enzymes. *Food Chem Toxicol* **2013**, *56*, 392–397.

(10) He, R. J.; Tian, Z. H.; Huang, J.; Sun, M. R.; Wei, F.; Li, C. Y.; Zeng, H. R.; Zhang, F.; Guan, X. Q.; Feng, Y.; Meng, X. M.; Yang, H.; Ge, G. B. Rationally Engineered CYP3A4 Fluorogenic Substrates for Functional Imaging Analysis and Drug-Drug Interaction Studies. *J Med Chem* **2023**, *66* (10), 6743–6755.

(11) Sliwoski, G.; Kothiwale, S.; Meiler, J.; Lowe, E. W. Computational Methods in Drug Discovery. *Pharmacol Rev* **2014**, *66* (1), 334–395.

(12) Ekroos, M.; Sjögren, T. Structural Basis for Ligand Promiscuity in Cytochrome P450 3A4. In *Proceedings of the National Academy of Sciences* **2006**, *103* (37), pp. 13682–13687.

(13) Scott, E. E.; Halpert, J. R. Structures of Cytochrome P450 3A4. *Trends Biochem Sci* **2005**, *30* (1), 5–7.

(14) Sun, H.; Veith, H.; Xia, M.; Austin, C. P.; Huang, R. Predictive Models for Cytochrome P450 Isozymes Based on Quantitative High Throughput Screening Data. *J Chem Inf Model* **2011**, *51* (10), 2474–2481.

(15) Plonka, W.; Stork, C.; Šícho, M.; Kirchmair, J. CYPlebrity: Machine Learning Models for the Prediction of Inhibitors of Cytochrome P450 Enzymes. *Bioorganic & Medicinal Chemistry* **2021**, *46*, 116388.

(16) Beck, T. C.; Beck, K. R.; Morningstar, J.; Benjamin, M. M.; Norris, R. A. Descriptors of Cytochrome Inhibitors and Useful Machine Learning Based Methods for the Design of Safer Drugs. *Pharmaceuticals* **2021**, *14* (5), 472.

(17) Silva, F.; Varlamova, E.; Braga, R.; Andrade, C. Development of QSAR Models for Identification of CYP3A4 Substrates and Inhibitors. In *Proceedings of MOL2NET, International Conference on Multidisciplinary Sciences*; MDPI: Sciforum.net, **2015**; pp. b034.

(18) Veith, H.; Southall, N.; Huang, R.; James, T.; Fayne, D.; Artemenko, N.; Shen, M.; Inglese, J.; Austin, C. P.; Lloyd, D. G.; Auld, D. S. Comprehensive Characterization of Cytochrome P450 Isozyme Selectivity across Chemical Libraries. *Nat Biotechnol* **2009**, *27* (11), 1050–1055.

(19) Kriegl, J. M.; Eriksson, L.; Arnhold, T.; Beck, B.; Johansson, E.; Fox, T. Multivariate Modeling of Cytochrome P450 3A4 Inhibition. *Eur J Pharm Sci* **2005**, *24* (5), 451–463.

(20) Li, X.; Xu, Y.; Lai, L.; Pei, J. Prediction of Human Cytochrome P450 Inhibition Using a Multitask Deep Autoencoder Neural Network. *Mol. Pharmaceutics* **2018**, *15* (10), 4336–4345.

(21) Yu, L.; Shi, X.; Tian, S.; Gao, S.; Li, L. Classification of Cytochrome P450 1A2 Inhibitors and Noninhibitors Based on Deep Belief Network. *Int. J. Comp. Intel. Appl.* **2017**, *16* (01), 1750002.

(22) Shao, C.-Y.; Su, B.-H.; Tu, Y.-S.; Lin, C.; Lin, O. A.; Tseng, Y. J. CypRules: A Rule-Based P450 Inhibition Prediction Server. *Bioinformatics* **2015**, *31* (11), 1869–1871.

(23) Zhang, S.; Yan, Z.; Huang, Y.; Liu, L.; He, D.; Wang, W.; Fang, X.; Zhang, X.; Wang, F.; Wu, H.; Wang, H. HelixADMET: A Robust and Endpoint Extensible ADMET System Incorporating Self-Supervised Knowledge Transfer. *Bioinformatics* **2022**, *38* (13), 3444–3453.

(24) Xiong, G.; Wu, Z.; Yi, J.; Fu, L.; Yang, Z.; Hsieh, C.; Yin, M.; Zeng, X.; Wu, C.; Lu, A.; Chen, X.; Hou, T.; Cao, D. ADMETlab 2.0: An Integrated Online Platform for Accurate and Comprehensive Predictions of ADMET Properties. *Nucleic Acids Research* **2021**, *49* (W1), W5–W14.

(25) Wei, Y.; Li, S.; Li, Z.; Wan, Z.; Lin, J. Interpretable-ADMET: A Web Service for ADMET Prediction and Optimization Based on Deep Neural Representation. *Bioinformatics* **2022**, *38* (10), 2863–2871.

(26) Venkatraman, V. FP-ADMET: A Compendium of Fingerprint-Based ADMET Prediction Models. *J Cheminform* **2021**, *13* (1), 75.

(27) Yang, H.; Lou, C.; Sun, L.; Li, J.; Cai, Y.; Wang, Z.; Li, W.; Liu, G.; Tang, Y. AdmetSAR 2.0: Web-Service for Prediction and Optimization of Chemical ADMET Properties. *Bioinformatics* **2019**, *35* (6), 1067–1069.

(28) Banerjee, P.; Dunkel, M.; Kemmler, E.; Preissner, R. SuperCYPsPred—a Web Server for the Prediction of Cytochrome Activity. *Nucleic Acids Research* **2020**, *48* (W1), W580–W585.

(29) Nguyen-Vo, T. H.; Trinh, Q. H.; Nguyen, L.; Nguyen-Hoang, P. U.; Nguyen, T. N.; Nguyen, D. T.; Nguyen, B. P.; Le, L. iCYP-MFE: Identifying Human Cytochrome P450 Inhibitors Using Multitask Learning and Molecular Fingerprint-Embedded Encoding. *J. Chem. Inf. Model.* **2022**, 62,21,5059-5068.

(30) Mervin, L. H.; Johansson, S.; Semenova, E.; Giblin, K. A.; Engkvist, O. Uncertainty Quantification in Drug Design. *Drug Discovery Today* **2021**, *26* (2), 474–489.

(31) Kendall, A.; Gal, Y. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? 2017, In *Advances in Neural Information Processing Systems*, **2017**, pp. 5574–5584.

(32) Abdar, M.; Pourpanah, F.; Hussain, S.; Rezazadegan, D.; Liu, L.; Ghavamzadeh, M.; Fieguth, P.; Cao, X.; Khosravi, A.; Acharya, U. R.; Makarenkov, V.; Nahavandi, S. A Review of Uncertainty Quantification in Deep Learning: Techniques, Applications and Challenges. *Information Fusion* **2021**, *76*, 243–297.

(33) Yu, J.; Wang, D.; Zheng, M. Uncertainty Quantification: Can We Trust Artificial Intelligence in Drug Discovery? *iScience* **2022**, *25* (8), 104814.

(34) Gal, Y.; Ghahramani, Z. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. arXiv October 4, 2016.

(35) Salakhutdinov, R.; Mnih, A*.* Bayesian Probabilistic Matrix Factorization Using Markov Chain Monte Carlo; In *Proceedings of the 25 th International Conference on Machine Learning*,Helsinki, Finland, **2008**.

(36) Blei, D. M.; Kucukelbir, A.; McAuliffe, J. D. Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association* **2017**, *112* (518), 859–877.

(37) Zhang, C.; Bütepage, J.; Kjellström, H.; Mandt, S. Advances in Variational Inference. In *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2019**, *41* (8), 2008–2026.

(38) Blundell, C.; Cornebise, J.; Kavukcuoglu, K.; Wierstra, D. Weight Uncertainty in Neural Networks. arXiv May 21, 2015.

(39) Stanton, S.; Maddox, W.; Delbridge, I.; Wilson, A. G. Kernel Interpolation for Scalable Online Gaussian Processes. arXiv March 3, 2015.

(40) Jain, S.; Liu, G.; Mueller, J.; Gifford, D. Maximizing Overall Diversity for Improved Uncertainty Estimates in Deep Ensembles. In *Proceedings of the AAAI Conference on Artificial Intelligence* **2020**, *34* (04), pp. 4264–4271.

(41) Duan, T.; Anand, A.; Ding, D. Y.; Thai, K. K.; Basu, S.; Ng, A.; Schuler, A. NGBoost: Natural Gradient Boosting for Probabilistic Prediction. In *Proceedings of the 37th International Conference on Machine Learning*; PMLR, **2020**; pp. 2690–2700.

(42) Breiman, L. Bagging Predictors. *Machine Learning* **1996**, *24* (2), 123–140.

(43) Maddox, W. J.; Garipov, T.; Izmailov, P.; Vetrov, D.; Wilson, A. Fast Uncertainty Estimates and Bayesian Model Averaging of DNNs. arXiv May 19, 2021.

(44) Sousa, M. Inductive Conformal Prediction: A Straightforward Introduction with Examples in Python. arXiv July 1, 2022.

(45) Vovk, V. Transductive Conformal Predictors. In *Artificial Intelligence Applications and Innovations*; AIAI 2013, IFIP AICT 412, **2013**; pp. 348–360.

(46) Amini, A.; Schwarting, W.; Soleimany, A.; Rus, D. Deep Evidential Regression. arXiv November 24, 2020.

(47) Sensoy, M.; Kaplan, L.; Kandemir, M. Evidential Deep Learning to Quantify Classification Uncertainty. arXiv October 31, 2018.

(48) Jensen, O.; Brockmöller, J.; Dücker, C. Identification of Novel High-Affinity Substrates of OCT1 Using Machine Learning-Guided Virtual Screening and Experimental Validation. *J. Med. Chem.* **2021**, *64* (5), 2762–2776.

(49) Mendolia, I.; Contino, S.; De Simone, G.; Perricone, U.; Pirrone, R. EMBER—Embedding Multiple Molecular Fingerprints for Virtual Screening. *International Journal of Molecular Sciences* **2022**, *23* (4), 2156.

(50) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50* (5), 742–754.

(51) Kukreja, S. L.; Löfberg, J.; Brenner, M. J. A LEAST ABSOLUTE SHRINKAGE AND SELECTION OPERATOR (LASSO) FOR NONLINEAR SYSTEM IDENTIFICATION. *IFAC Proceedings Volumes* **2006**, *39* (1), pp. 814–819.

(52) Zhang, F.; Liu, W.; Huang, J.; Chen, Q.; Wang, D.; Zou, L.; Zhao, Y.; Zhang, W.; Xu, J.; Chen, H.; Ge, G. Inhibition of Drug-Metabolizing Enzymes by Jingyin Granules: Implications of Herb–Drug Interactions in Antiviral Therapy. *Acta Pharmacol Sin* **2022**, *43* (4), 1072–1081.

(53) Ning, J.; Wang, W.; Ge, G.; Chu, P.; Long, F.; Yang, Y.; Peng, Y.; Feng, L.; Ma, X.; James, T. D. Target Enzyme‐Activated Two‐Photon Fluorescent Probes: A Case Study of CYP3A4 Using a Two‐Dimensional Design Strategy. *Angew. Chem. Int. Ed.* **2019**, *58* (29), 9959–9963.

(54) Wang, D.; Wu, Z.; Shen, C.; Bao, L.; Luo, H.; Wang, Z.; Yao, H.; Kong, D.-X.; Luo, C.; Hou, T. Learning with Uncertainty to Accelerate the Discovery of Histone Lysine-Specific Demethylase 1A (KDM1A/LSD1) Inhibitors. *Briefings in Bioinformatics* **2022**, bbac592.

(55) Kuleshov, V.; Fenner, N.; Ermon, S. !!46 Accurate Uncertainties for Deep Learning Using Calibrated Regression. In *Proceedings of the 35th International Conference on Machine Learning*; PMLR, **2018**; pp. 2796–2804.