

RESEARCH

Open Access



GTransCYPs: an improved graph transformer neural network with attention pooling for reliably predicting CYP450 inhibitors

Candra Zonyfar¹, Soualihou Ngnamsie Njimbouom¹, Sophia Mosalla² and Jeong-Dong Kim^{1,2,3*}

Abstract

State-of-the-art medical studies proved that predicting CYP450 enzyme inhibitors is beneficial in the early stage of drug discovery. However, accurate machine learning-based (ML) *in silico* methods for predicting CYP450 inhibitors remains challenging. Here, we introduce GTransCYPs, an improved graph neural network (GNN) with a transformer mechanism for predicting CYP450 inhibitors. This model significantly enhances the discrimination between inhibitors and non-inhibitors for five major CYP450 isozymes: 1A2, 2C9, 2C19, 2D6, and 3A4. GTransCYPs learns information patterns from molecular graphs by aggregating node and edge representations using a transformer. The GTransCYPs model utilizes transformer convolution layers to process features, followed by a global attention-pooling technique to synthesize the graph-level information. This information is then fed through successive linear layers for final output generation. Experimental results demonstrate that the GTransCYPs model achieved high performance, outperforming other state-of-the-art methods in CYP450 prediction.

Scientific contribution

The prediction of CYP450 inhibition via computational techniques utilizing biological information has emerged as a cost-effective and highly efficient approach. Here, we presented a deep learning (DL) architecture based on GNN with transformer mechanism and attention pooling (GTransCYPs) to predict CYP450 inhibitors. Four GTransCYPs of different pooling technique were tested on an experimental tasks on the CYP450 prediction problem for the first time. Graph transformer with attention pooling algorithm achieved the best performances. Comparative and ablation experiments provide evidence of the efficacy of our proposed method in predicting CYP450 inhibitors. The source code is publicly available at <https://github.com/zonwoo/GTransCYPs>.

Keywords Cytochrome P450 inhibition, Drug-drug interaction, Deep learning, Graph transformer neural network, Attention mechanism, Graph pooling

Introduction

In drug-drug interactions, the inhibition of cytochrome P450 (CYP450) enzymes plays a crucial role in drug efficacy, toxicity, and potential interactions [1–4]. These enzymes are responsible for metabolizing numerous drugs in the body [5]. If the activity of these enzymes is hindered by one drug, it can impact the metabolism of other drugs, potentially altering the drug's response and raising the risk of toxicity. Among the 57 commonly found CYP450 isozymes in the human liver [2, 3, 6], five

*Correspondence:

Jeong-Dong Kim

kjd4u@sunmoon.ac.kr

¹ Department of Computer Science and Electronic Engineering, Sun Moon University, Asan 31460, Republic of Korea

² Division of Computer Science and Engineering, Sun Moon University, Asan 31460, Republic of Korea

³ Genome-based BioIT Convergence Institute, Sun Moon University, Asan 31460, Republic of Korea



© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

of them—namely 1A2, 2C9, 2C19, 2D6, and 3A4—play critical roles in most drug metabolism processes in the human body [7].

In vitro, high-throughput screening for CYP450 inhibition technology has generated data on CYP450 isozymes, including through research initiatives such as PubChem Bioassays [8]. The collected data has enabled the computational prediction of potential inhibitory compounds against five CYP450 isozymes using ML approaches. The in silico approach is appealing as it can be utilized at the early stages of drug discovery pathways, reducing the number of wet-lab experiment studies needed for selecting new drug candidates and thus minimizing costs. In addition to enhancing success rates, it also aids in predicting the activity of designed compounds before synthesis [9, 10].

In recent years, ML has been used as a computational method to predict CYP450 inhibition [11–16]. Various ML algorithms have been applied in this research, including random forest (RF) [15], and support vector machines (SVM) [17]. RF, which is a collection of decision trees (DT), is used to improve accuracy and reduce overfitting by combining the results from several DT. Additionally, SVM is used to build a model that can separate data into different classes by finding the optimal hyperplane in feature space.

In the biomedical domain and molecular property prediction, GNN are attracting growing interest and have now set state-of-the-art methods [18–21]. Unlike conventional ML models commonly employed in research to predict CYP450 enzyme activity relying on chemical features and molecular structure, GNN strives to utilize data representation in graph form. GNN facilitates the integration of topological information from molecular graph structures into the model, thereby considering the spatial relationships among atoms within molecules. Its ability to model complex molecular structures and account for potential atom interactions can enhance prediction accuracy. A model based on GNN has been proposed by Qiu et al. [22] to predict CYP450 inhibition. They utilized two types of input data for the developed model. On one side, they extracted chemical representation features from SMILES into the GNN, while on the other side, they extracted features from sequence alignments with convolutional neural network (CNN). Subsequently, these two sets of features were concatenated at the end of the model. As a result, their model was reported to outperform the iCYP-MFE [16] model. On the other hand, recent research in predicting CYP450 activity has been conducted by Ai et al. [23]. The method they proposed involves two pathways. The first pathway employs an artificial neural network (ANN) to learn information from substructure-based molecular fingerprints, as well as one

pharmacophore-based fingerprint. The second pathway utilizes a GNN with attention mechanisms to extract structural information from molecular graphs, which is then combined with the features obtained from the first pathway in a fully connected layer. Their proposed model, named FPGNN, integrates these pathways to predict the inhibition of five CYP450 isozymes. However, despite promising results, incorporating fingerprints and GNN presents a limitation. Firstly, there is a tendency towards information duplication, as fingerprint features may already encapsulate data that GNN could learn from the molecular structure. This redundancy can hinder the model's capacity to discern genuinely informative features. Secondly, merging various feature representations increases computational complexity, potentially hindering the model's effective prediction.

In this paper, we proposed a DL model, an improved GNN with a transformer mechanism for predicting CYP450 inhibitors (GTransCYPs). Initially, the drug chemical structure is represented as graph in which the vertices are atoms, and the edges are chemical relationships within the molecule. Next, a graph transformer network is used to compute the drug embedding vectors. In addition, attention pooling is used to enable more efficient downsampling, improving model generalizability. Results from extensive experiments demonstrate that GTransCYPs improves the performance of predicting CYP450 inhibitors in comparison with the state-of-the-art models. An ablation study was carried out to show the robustness of the proposed method in modeling parameters and their ability to predict potential inhibitory compounds against five CYP450 isozyme. In summary, we believed that GTransCYPs is an effective tool to identify potential inhibitory compounds against CYP450 for further wet-lab experiment validation. A web server was developed to host the models from the study for public access.

Method

Overview of the GTransCYPs model

Graph learning models garner growing attention in the computational analysis of molecule data. The proposed model is instantiated through GNN class, designed to extract molecular data and encapsulate insights from atomic node characteristics by orchestrating a stratified transformation process. This network efficiently acquires hierarchically enriched node embeddings, considering local neighbor interactions and comprehensive graph comprehension. Key components of this network architecture include the integration of transformer layers (TransformerConv), linear projection, batch normalization, and attention pooling, as illustrated in Fig. 1. These elements underscore contextual awareness's significance

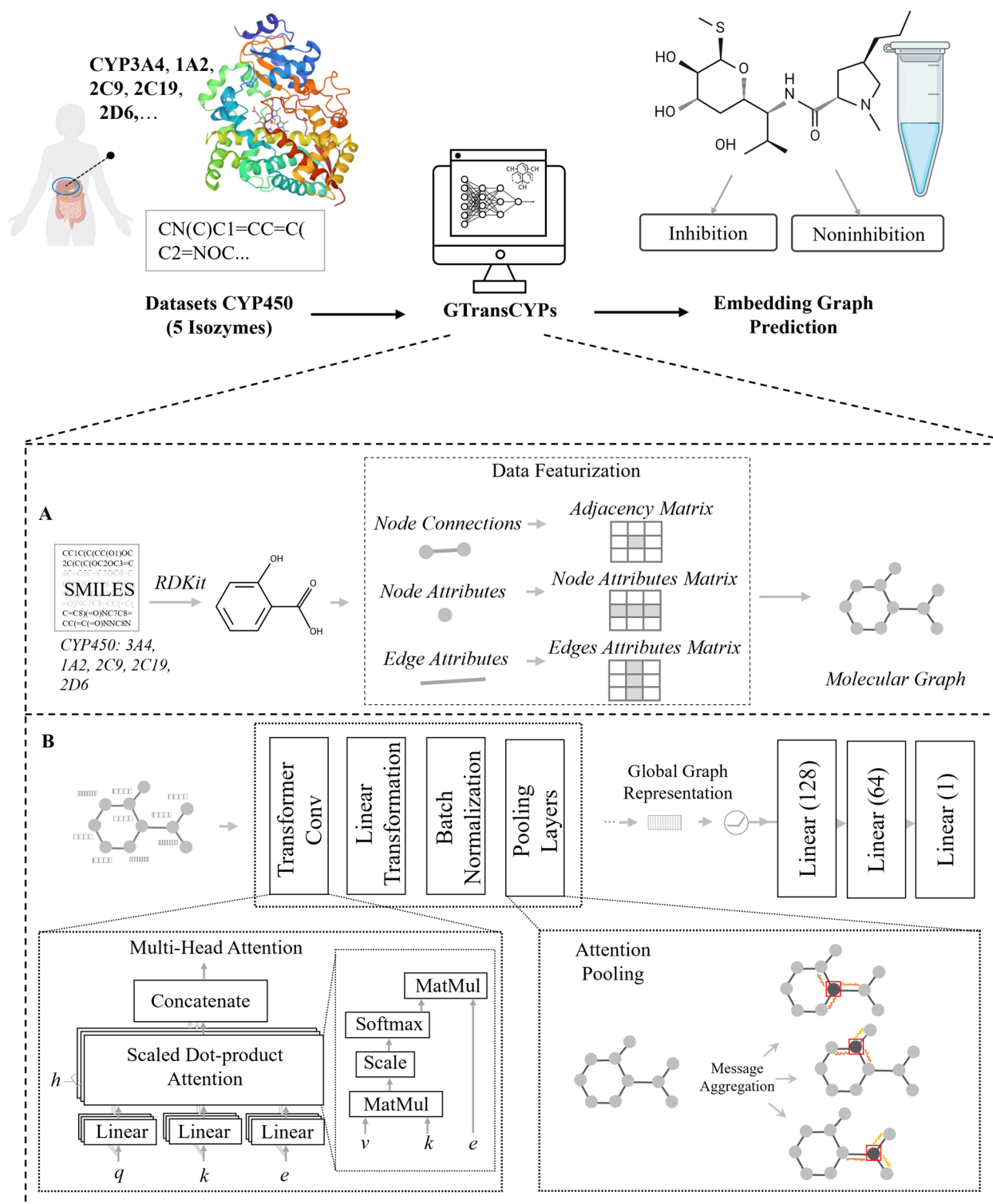


Fig. 1 An overview of GTransCYP5

in local and global contexts. It aims to empower networks to learn complex graph representations. Thus, finding a harmonic balance between capturing delicate graph patterns and maintaining a thorough understanding

will improve the predictive performance of CYP450 inhibitors.

Figure 1A shows an overview of the molecular graph construction, the pertinent features are extracted from

the data of the five CYP450 isozymes through a featurization process. In this method, each molecule is transformed into a graph wherein the atoms serve as the nodes and the interconnecting bonds between the atoms are depicted as edges in the form of numerical vectors. To accomplish this transformation, we make use of the RDKit library, which facilitates the conversion of the SMILES representation into the requisite structural format for training the model. RDKit is used for its capability to manipulate molecular structures, enabling comprehensive representation and processing of molecular data. Figure 1B shows the model initiates the process by transforming atomic input into an initial embedding representation utilizing multi-head attention, succeeded by normalization. Subsequently, aggregation is performed through graph pooling operations, specifically attention pooling layers. Following this, the data traverse multiple linear layers and undergoes activation through the Rectified Linear Unit (ReLU) function to produce the final output.

Graph neural network (GNN)

GNN has become increasingly popular and reliable in molecular analysis due to their ability to capture high-level graph information and relationships and propagate them through networks. GNN is a special type of DL technique created explicitly for processing and analyzing structured data in the form of graphs [24]. The development of GNN has adopted the concept of attention mechanisms and allows models to assign dynamic weights to interactions between nodes in a graph, enabling more adaptive decision making and capturing long-range dependencies. GNN can be applied to the graph representation of compound molecule (G) by using an iterative mechanism to update node and edge features based on their neighbors. This process can be described in the following equations:

$$h_v^{(l+1)} = \phi_v \left(h_v^{(l)}, h_e^{(l)} | (v_{a_i}, v_{a_j}) \in E \right) \quad (1)$$

$$h_e^{(l+1)} = \phi_e \left(h_e^{(l)}, \left\{ h_v^{(l+1)} | v_a \text{ is neighbor}(v_{a_i}, v_{a_j}) \right\} \right) \quad (2)$$

where $h_v^{(l)}$ is the representation of the v_a node feature on the t -th iteration, and $h_e^{(l)}$ is the representation of the edge feature (v_{a_i}, v_{a_j}) on the l -th iteration.

Molecular graph construction

The graph concept enables encoding the high-dimensional space of molecular structures into a lower-dimensional representation. The simplified molecular input line entry

system (SMILES) of CYP450 enzyme was introduced as a two-dimension molecular graph, representing chemical atoms and bond token information. The molecular graph of the CYP450 is represented as $G = (V, E)$ where V denotes the set of nodes and E indicates the edges. Each $a \in A$ atom is mapped to a node in the graph, so $V = \{v_a | a \in A\}$, where v_a is the node representing the a atom. Furthermore, each $b \in B$ bond is mapped to an edge between nodes representing the atoms bound by that bond, therefore $E = \{(v_{a_i}, v_{a_j}) | b = (a_i, a_j) \in B\}$. In addition, features are added at each node and edge to represent atomic and bonding properties. Suppose F_v is the feature space for nodes and F_e is the feature space for edges. The features on the v_a node can represent the atomic type, charge, and other properties of the E atom, $f(v_a) = (F_{atom}(a), F_{charge}(a), \dots)$. On the other hand, features on the edge (v_{a_i}, v_{a_j}) can represent the bond type, bond length, or other properties of bonds $b = (a_i, a_j)$, that is $f(v_{a_i}, v_{a_j}) = (F_{bond}(b), F_{length}(b), \dots)$.

We use molecular featurization to handle graph construction by adopting featurization based on path-augmented graph transformer network [25], a special feature that builds a molecular graph connecting all pairs of atoms that takes into account the interaction of atoms with every other atom in the molecule. Path-augmented graph transformer networks were employed in the featurization process to capture both local and global structural information of molecules, providing a context-aware representation of molecular features. The featurization process is carried out iteratively on each molecular entity in the dataset. For each molecule, its SMILES representation is converted to a molecular graph object using the help of DeepChem [26] and the RDKit library [27]. The graph is then converted into a feature representation, including node attributes and edge attributes.

Graph transformer for learning CYP450 molecules

The proposed model is designed to understand the representation of molecular convolutional features in graph structures, considering the information of each atom in the molecule. With the node features provided $H^l = \{h_1^{(l)}, h_2^{(l)}, \dots, h_n^{(l)}\}$, the multi-head attention for each edge is computed as follows:

$$q_{c,i}^{(l)} = W_{c,q}^{(l)} h_i^{(l)} + b_{c,q}^{(l)} \quad (3)$$

$$k_{c,j}^{(l)} = W_{c,k}^{(l)} h_j^{(l)} + b_{c,k}^{(l)} \quad (4)$$

$$e_{c,ij} = W_{c,e} e_{ij} + b_{c,e} \quad (5)$$

$$a_{c,ij}^{(l)} = \frac{\langle q_{c,i}^{(l)}, k_{c,j}^{(l)}, e_{c,ij} \rangle}{\sum_{u \in N(i)} \langle q_{c,i}^{(l)}, k_{c,u}^{(l)}, e_{c,iu} \rangle} \quad (6)$$

Here, q , k represent the query, and key projections, respectively, at layer l . These projections are computed using weight $W_{c,q}^{(l)}$, $W_{c,k}^{(l)}$ and biases $b_{c,q}^{(l)}$, $b_{c,k}^{(l)}$. They are linear transformations of the original input tokens' representations $h_i^{(l)}$ and $h_j^{(l)}$. Then, $e_{c,ij}$ quantifies the importance or score of the interaction between i -th query and j -th key. It is computed by applying another linear transformation with weight $W_{c,e}$ and bias $b_{c,e}$. $a_{c,ij}^{(l)}$ is the attention weight associated with i -th query and j -th key at layer l . It is computed by normalizing the dot product of the $q_{c,i}^{(l)}$, $k_{c,j}^{(l)}$, $e_{c,ij}$. The denominator involves a sum over all attention scores between the i -th query and other keys, represented by $k_{c,u}^{(l)}$ for u in the neighborhood $N(i)$.

In this context, each token in an order that represents the molecular structure maps to a node in the graph. Key projections $k_{c,j}^{(l)}$ and query projections $q_{c,i}^{(l)}$ are calculated by multiplying the corresponding weight $W_{c,q}^{(l)}$, $W_{c,k}^{(l)}$ by the representation of each atoms or tokens in a molecule, which serve as key-value and query-value respectively. Then, the edge features $e_{c,ij}$ are calculated by performing the inner product between the key projection and the query projection for each pair of atoms. Weights $W_{c,e}$ and $b_{c,e}$ and exponential transformations play an important role in giving emphasis or increasing the attention score according to their relevance to molecular interactions, this allows the model to explore complex interactions between atoms in molecules. Furthermore, $a_{c,ij}^{(l)}$ is used to calculate the attention weight which indicates how important the interaction between i and j atoms is in informing enzyme activity. In a nutshell, this process allows the model to adaptively understand the complexity of molecular structure and the interrelationships of its atoms, contributing to the model's ability to make predictions.

Message aggregation with transformer

In this study, we employ message passing mechanism through neural network to get local information in the CYP450 molecular graph, then the environmental information of each node is aggregated to update the representation of the central node. Here, following the acquisition of multi-head attention from the graph, a process of message aggregation with transformer is executed:

$$v_{c,i}^{(l)} = W_{c,v}^{(l)} h_j^{(l)} + b_{c,v}^{(l)} \quad (7)$$

$$\widehat{h}_i^{(l+1)} = \parallel_{c=1}^C \left[\sum_{j \in N} a_{c,ij}^{(l)} (v_{c,i}^{(l)} + e_{c,ij}) \right] \quad (8)$$

where $\widehat{h}_i^{(l+1)}$ represents the updated representation of element i in the next layer $l + 1$, and where the \parallel is the concatenation operation for C head attention. This process involves aggregating messages originating from neighboring nodes. The contribution of each neighboring nodes, which is determined by $a_{c,ij}^{(l)}$, is calculated using the information contained in $v_{c,i}^{(l)}$ and the terms of edge features.

Attention pooling

In order to reduce the size and complexity of graph data structures while preserving significant information. In this study, we employ the attention pooling technique to streamline the graph representation while retaining essential information, thereby improving the model's performance. The attention mechanism empowers the model to prioritize crucial input elements necessary for task completion, thus enhancing its ability for optimal performance. Utilizing the attention technique, an examination is conducted on a sequence of normalized weights that reflect the relative significant levels attributed to each node. The attention is calculated as follows:

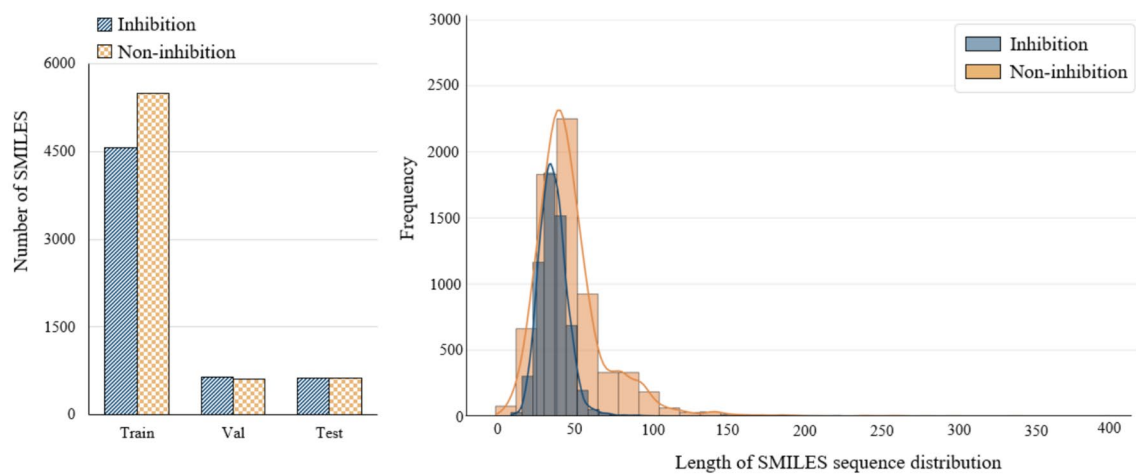
$$Attention_i = SoftMax_i(H \times W) = \frac{\exp(H_i \times W)}{\sum_{j \in N} \exp(H_j \times W)} \quad (9)$$

Here, the SoftMax function is used to normalize the attention vector, ensuring that the attention coefficients are proportional across various nodes.

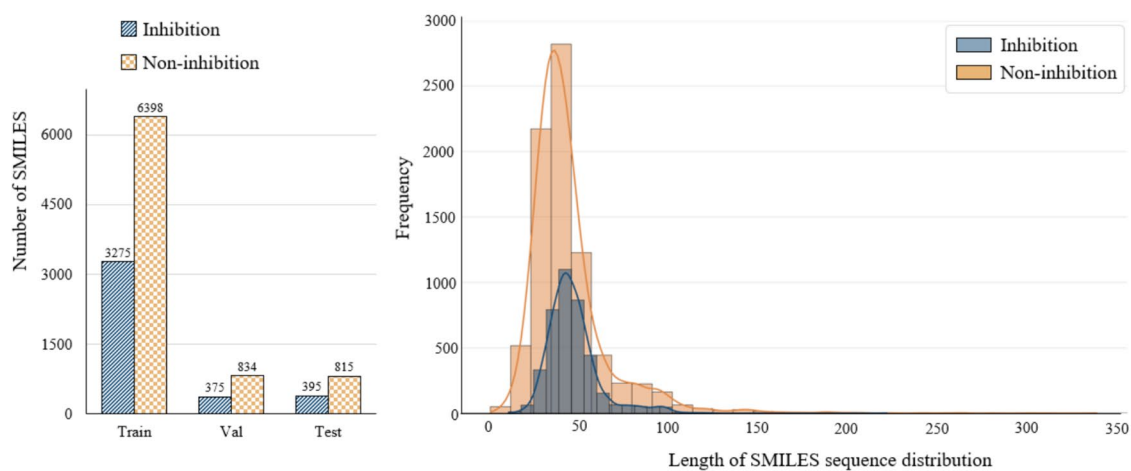
Experimental results

Datasets

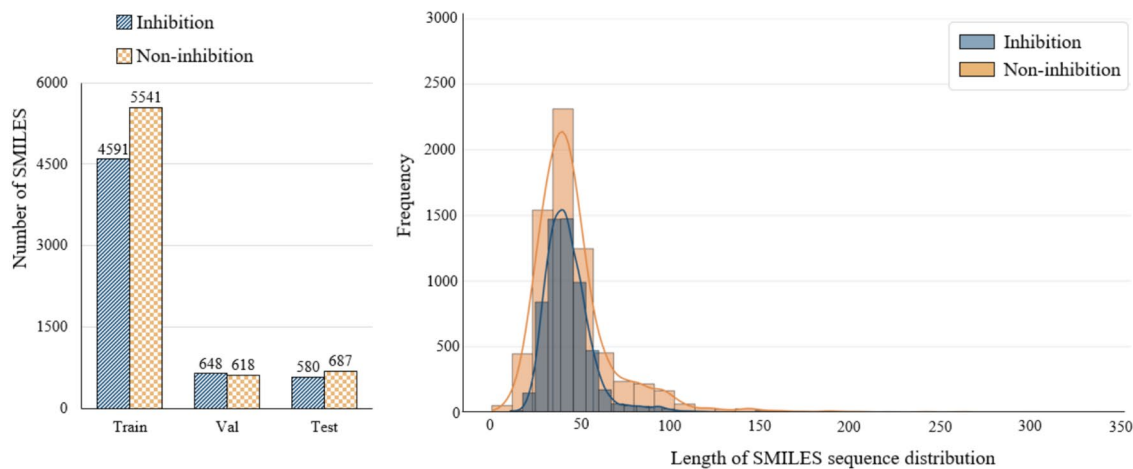
To validate the performance of our proposed model, we utilize the same dataset that was employed by Veith et al. [28] and collect from therapeutics data commons database [29] (<https://tdcommons.ai>, accessed January 2024). The dataset includes inhibitors targeting the five major CYP450 isozymes, namely 1A2, 2C9, C19, 2D6, and 3A4. We apply the scaffolding method to split the training, validation, and testing sets in an 80:10:10 ratio. This scaffolding method, which separates samples based on their two-dimensional structural frameworks, was chosen because it presents a greater challenge for learning algorithms compared to random splitting, while ensuring a better representation of molecular diversity in each subset. The training set comprises 4564 inhibitors and 5499 non-inhibitors for 1A2; 3275 inhibitors and 6398 non-inhibitors for 2C9; 4591 inhibitors and 5541



A) 1A2

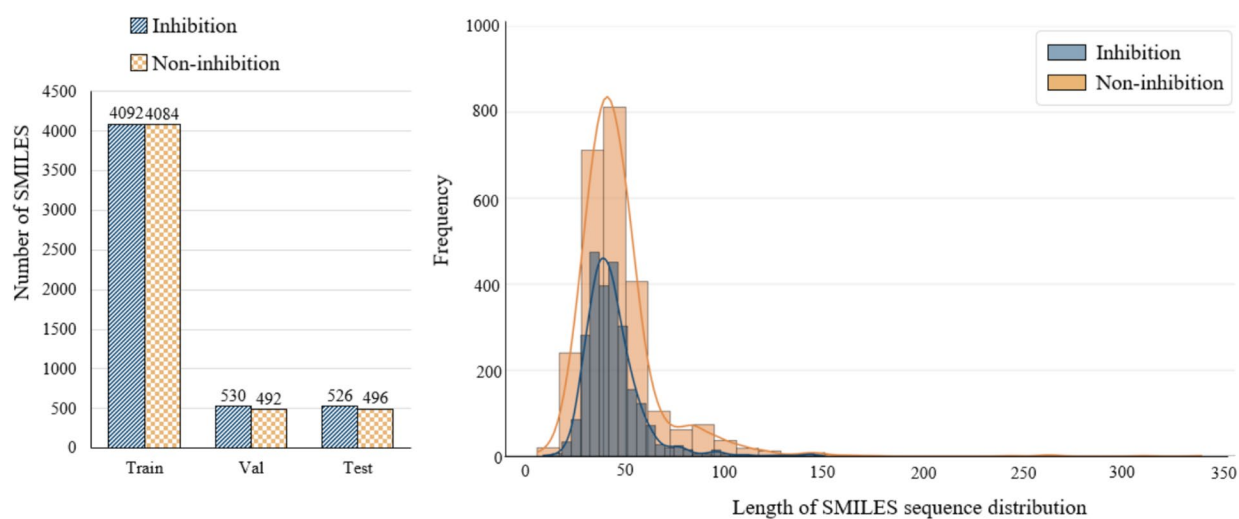


B) 2C9

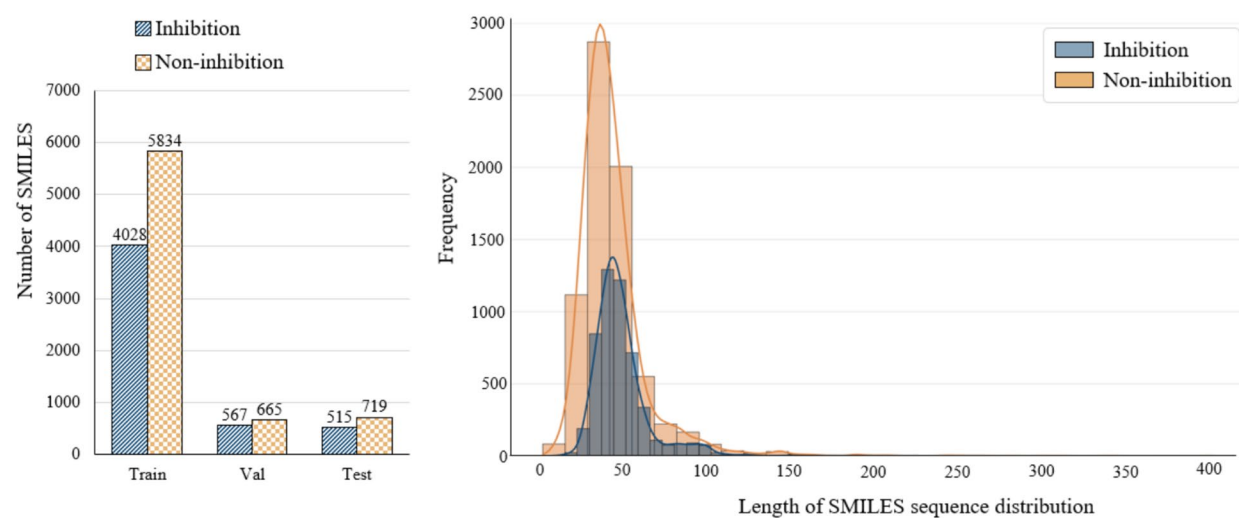


C) 2C19

Fig. 2 Dataset details for five CYP450 isozymes. Number of inhibition and non-inhibition data in training, validation, and testing sets, along with analysis of SMILES length distribution for (A) 1A2, (B) 2C9, (C) 2C19, (D) 2D6, and (E) 3A4 datasets



D) 2D6



E) 3A4

Fig. 2 continued

non-inhibitors for 2C19; and 4028 inhibitors and 5834 non-inhibitors for 3A4. The validation set contains 639 inhibitors and 618 non-inhibitors for 1A2; 375 inhibitors and 834 non-inhibitors for 2C9; 648 inhibitors and 618 non-inhibitors for 2C19; along with 567 inhibitors and 665 non-inhibitors for 3A4. Finally, the testing set comprises 626 inhibitors and 633 non-inhibitors for 1A2; 395 inhibitors and 815 non-inhibitors for 2C9; 580 inhibitors and 687 non-inhibitors for 2C19; and 515 inhibitors and 719 non-inhibitors for 3A4. Notably, the 2D6 dataset shows a significant imbalance with only 5148 instances labeled as inhibitors compared to 10,616 non-inhibitors,

indicating significant imbalance compared to datasets 2C9, 2C19, and 3A4. We implemented a downsampling strategy to balance the class distribution and address the disparity between the two classes of molecules in the 2D6 dataset. Oversampling, in contrast to downsampling, was not used because it addresses class imbalance by generating additional data for the minority class. However, this synthesized data often does not accurately represent real-world conditions. Initially, the dataset was partitioned by labels to isolate the majority (non-inhibitors) and minority (inhibitors) classes. We then applied downsampling to the non-inhibition data, randomly selecting samples without replacement to match the number of

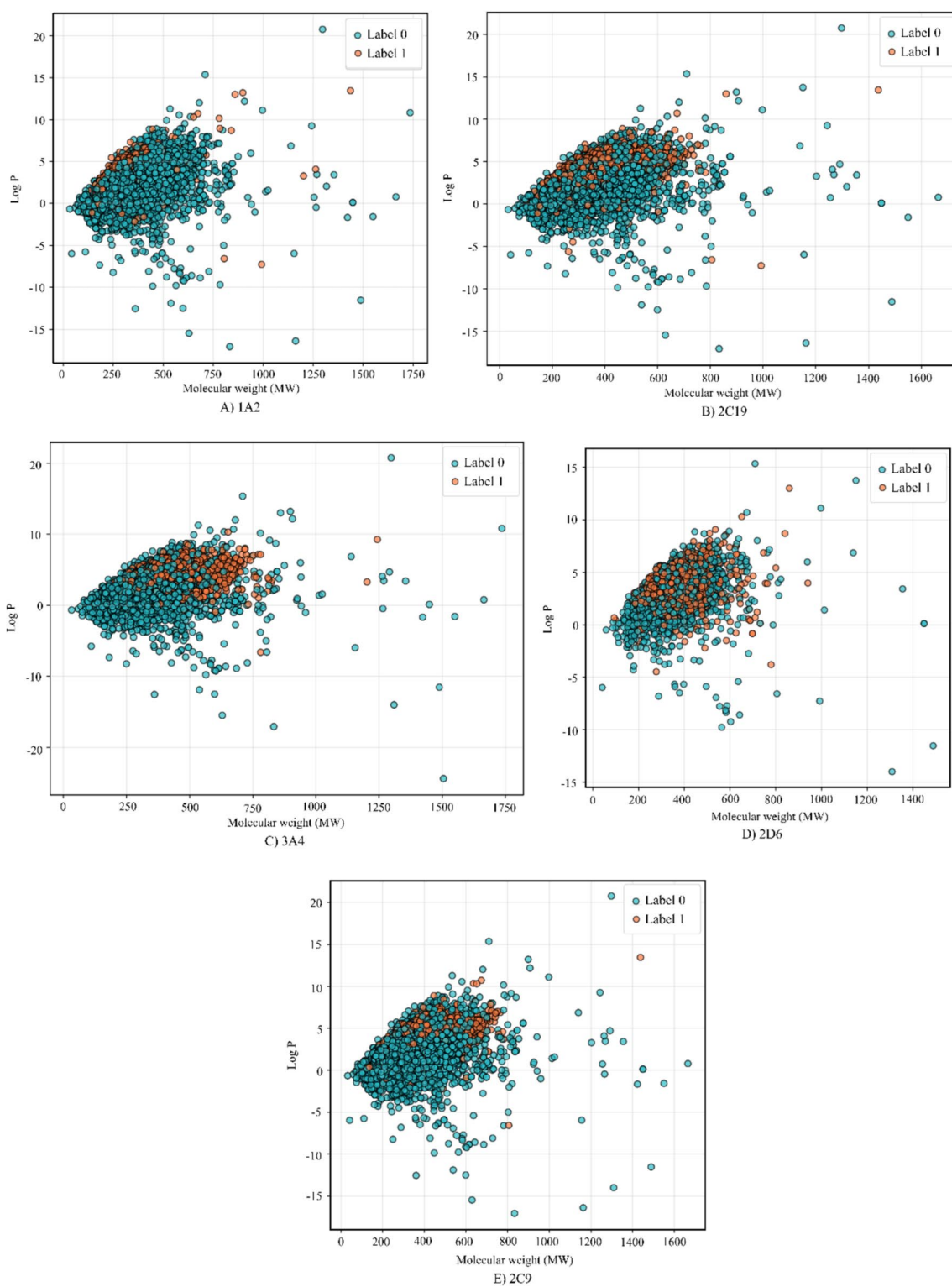


Fig. 3 The molecular chemical space distribution of CYP450 (1A2, 2C19, 3A4, 2D6, and 2C9) are illustrated. Chemical space is characterized using $\text{Log}P$ and molecular weight

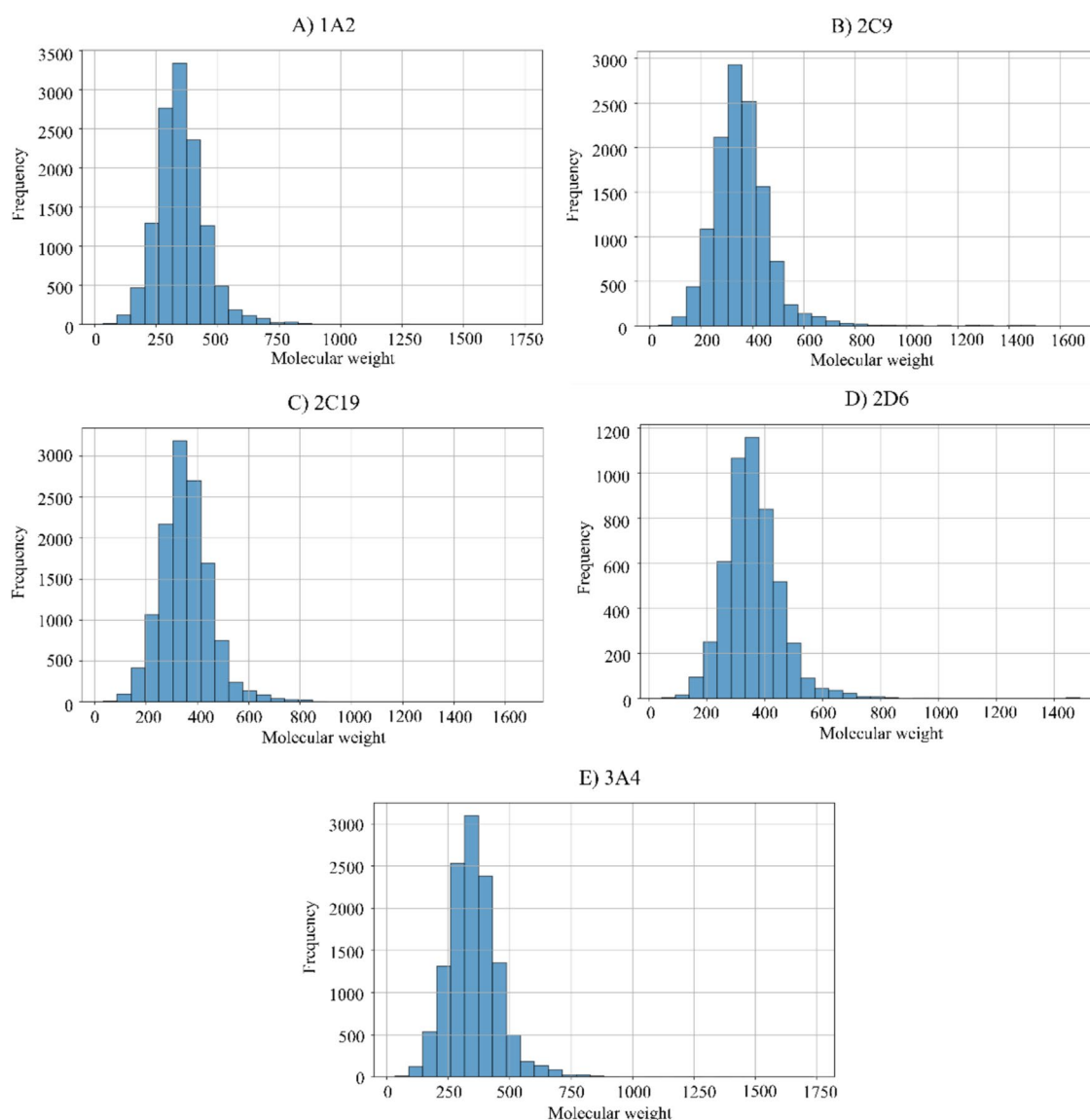


Fig. 4 Molecular weight distribution of each CYP450 isozyme

inhibition-labeled instances. This downsampled non-inhibition data was subsequently reintegrated with the inhibition data, achieving a more balanced class distribution to mitigate potential biases arising from the initial class imbalance. Furthermore, a dataset was divided into training, validation, and testing sets using scaffolding split technique. As a result, the training set involved 4092 inhibitors and 4084 non-inhibitors, validation set 530 inhibitors and 492 non-inhibitors; testing set 526 inhibitor and 496 non-inhibitors. Figure 2 presents a comprehensive analysis of inhibitor and non-inhibitor quantities within the training, validation, and testing sets. Additionally, it includes graphical representations depicting

the distribution of SMILES lengths across each dataset. SMILES length refers to the number of characters in the SMILES string that represents a molecule.

Figure 3 shows the chemical space represented by each dataset ascertained and compared, depending on related molecular descriptors such as molecular weight (MW) and $\text{Log}P$, adopted from [11, 23]. CYP450 inhibitors are visually represented by orange dots, while cyan dots correspond to CYP450 non-inhibitors. The compounds within the CYP450 modeling dataset exhibit a wide distribution across MW, reflecting a diverse array of chemical structures. This spectrum of MW and $\text{Log}P$ values underscore the dataset's inclusion of compounds spanning a broad

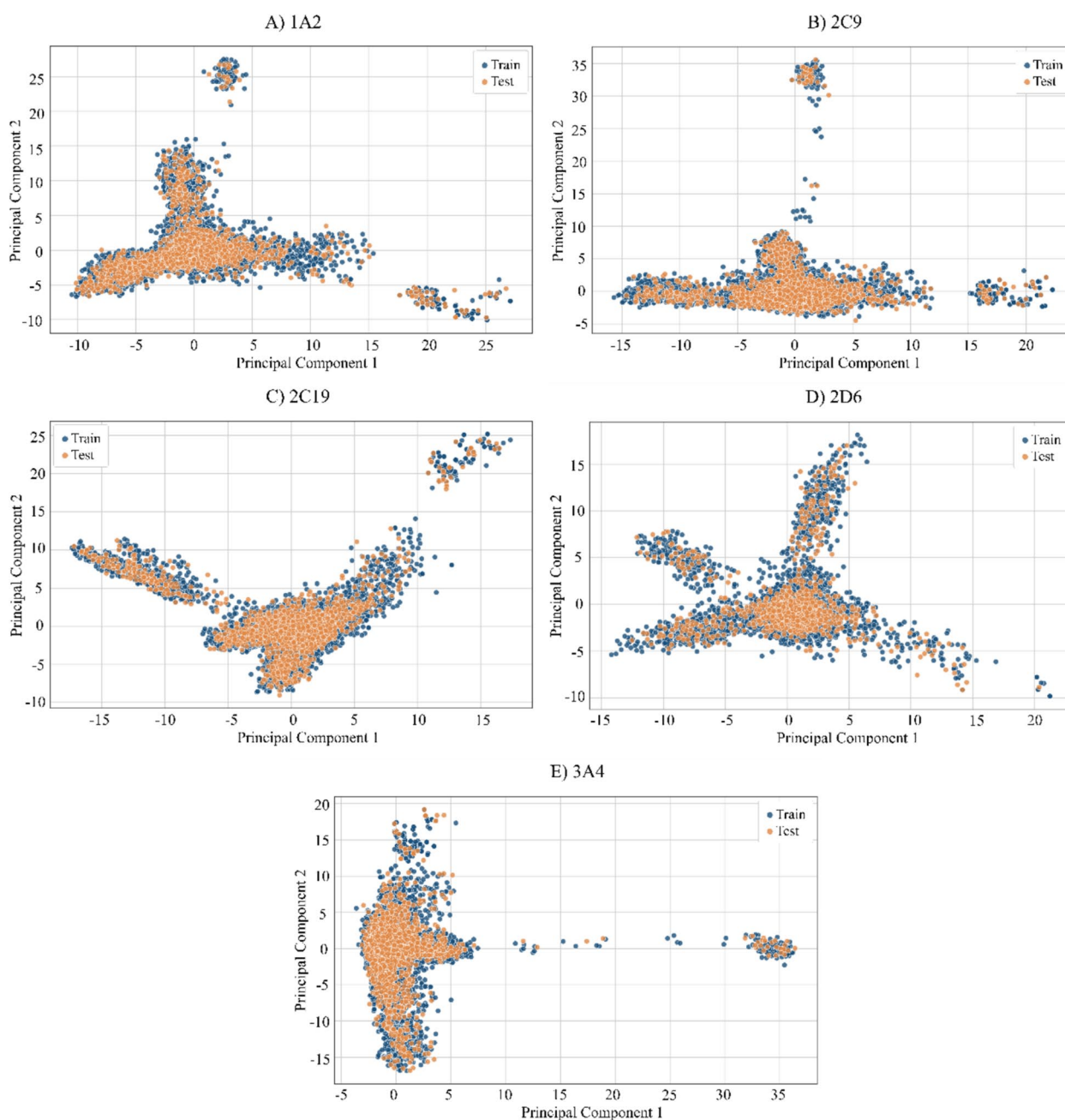


Fig. 5 PCA visualization of chemical space distribution of each CYP450 isozyms

chemical range. The chemical spaces of various CYP450 enzymes, including CYP1A2 (MW: 33.03 to 1736.18, LogP : -17.08 to 20.75), CYP2C9 (MW: 33.03 to 1664.92, LogP : -17.08 to 20.75), CYP2C19 (MW: 33.03 to 1664.92, LogP : -17.08 to 20.75), CYP2D6 (MW: 42.39 to 1488.80, LogP : -14.01 to 15.34), and CYP3A4 (MW: 33.03 to 1736.18, LogP : -24.39 to 20.75), highlight the encompassing variety of chemical characteristics in the dataset. In addition, we

analyzed the molecular weight distribution of compounds within the dataset, as shown in Fig. 4. For 1A2, the molecular weight of compounds ranged from 33.03 Dalton to 1736.18 Dalton, with a distribution peak at 291.35 Dalton. Datasets 2C9 and 2C19 showed similar molecular weight ranges, from 33.03 Dalton to 1664.92 Dalton, with modes at 280.33 Dalton and 291.35 Dalton, respectively. Dataset 2D6 had a molecular weight range from 42.39 Dalton to 1488.81 Dalton, with a mode at 291.35 Dalton. Lastly, 3A4

Table 1 The hyperparameters

Parameters	Range
Batch size	32, 64, 100, 128, 256
Learning rate	0.1, 0.01, 0.001, 0.0001
Weight decay	0.0001, 0.00001, 0.001
SGD momentum	0.9, 0.8, 0.5
Scheduler gamma	0.995, 0.9, 0.8, 0.5, 1
Pos weight	1.3
Embedding size	16, 32, 64, 128
Attention head	1, 2, 4, 8, 16
Model layers	1, 4, 8, 16, 32
Dropout rate	0.2, 0.5
Top K ratio	0.5
Top K every N	1
Dense neurons	32, 64, 128, 256

Table 2 Environmental setup

Environments	Descriptions
RAM	62.7 GB
CPU	Intel Core i9-9900 k (3.60 GHz)
GPU	NVIDIA GTX 1080 Ti×4
OS	Ubuntu 18.04.64bit
Python Version	3.8.0
Pytorch	PyTorch 2.0.1 + cu117
Pytorch-geometric	2.3.1
Pytorch-lightning	2.0.2
Deepchem	2.7.1
RDKit	2023.03.1
Sklearn	1.2.2
Pandas	2.0.2
Seaborn	0.12.2
CUDA Version	11.2.0

Table 3 Performance results of GTransCYPs on five isozyme CYP450 datasets

Datasets	1A2	2C9	2C19	2D6	3A4
BA	0.886	0.873	0.795	0.805	0.770
MCC	0.770	0.746	0.591	0.612	0.534
REC	0.916	0.912	0.846	0.868	0.823
PRE	0.924	0.911	0.744	0.748	0.677
F1-score	0.920	0.911	0.792	0.803	0.743

Table 4 Performance comparison of utilizing different pooling combinations in the ablation experiment

Models	BA	MCC	F1-score
model-A	0.836	0.671	0.836
model-B	0.879	0.755	0.879
model-C	0.875	0.751	0.875
model-D	0.886	0.770	0.886

* Bold denotes the highest value

displayed a molecular weight range from 33.03 Dalton to 1736.18 Dalton, with a mode at 291.35 Dalton. This analysis indicates a broad molecular weight distribution across the datasets, with a consistent mode around 291.35 Dalton for most datasets. Principal Component Analysis (PCA) was plotted to visualize and evaluate the chemical space coverage of the training and test datasets of each of the five CYP450 inhibitors. Figure 5 shows a scatterplot of the first two principal components, illustrating a similar distribution of chemical space between the training and test sets. This similarity helps define the applicability domain of the developed model, ensuring reliable assessment of new compounds chemically comparable to those in the training set. Ensuring that the test set is representative of the training set allows for an accurate evaluation of the model's performance and avoids potential biases from dissimilar chemical spaces.

Evaluation metrics

GTransCYPs model is evaluated using metrics such as balanced accuracy (BA), matthews correlation coefficient (MCC), precision (PRE), recall (REC), F1-Score. The calculations for these metrics are as follows:

$$BA = \frac{\frac{TP}{TP+FN} + \frac{TN}{FP+TN}}{2} \quad (10)$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) + (TP + FN) + (TN + FP) + (TN + FN)}} \quad (11)$$

$$PRE = \frac{TP}{TP + FP} \quad (12)$$

$$REC = \frac{TP}{TP + FN} \quad (13)$$

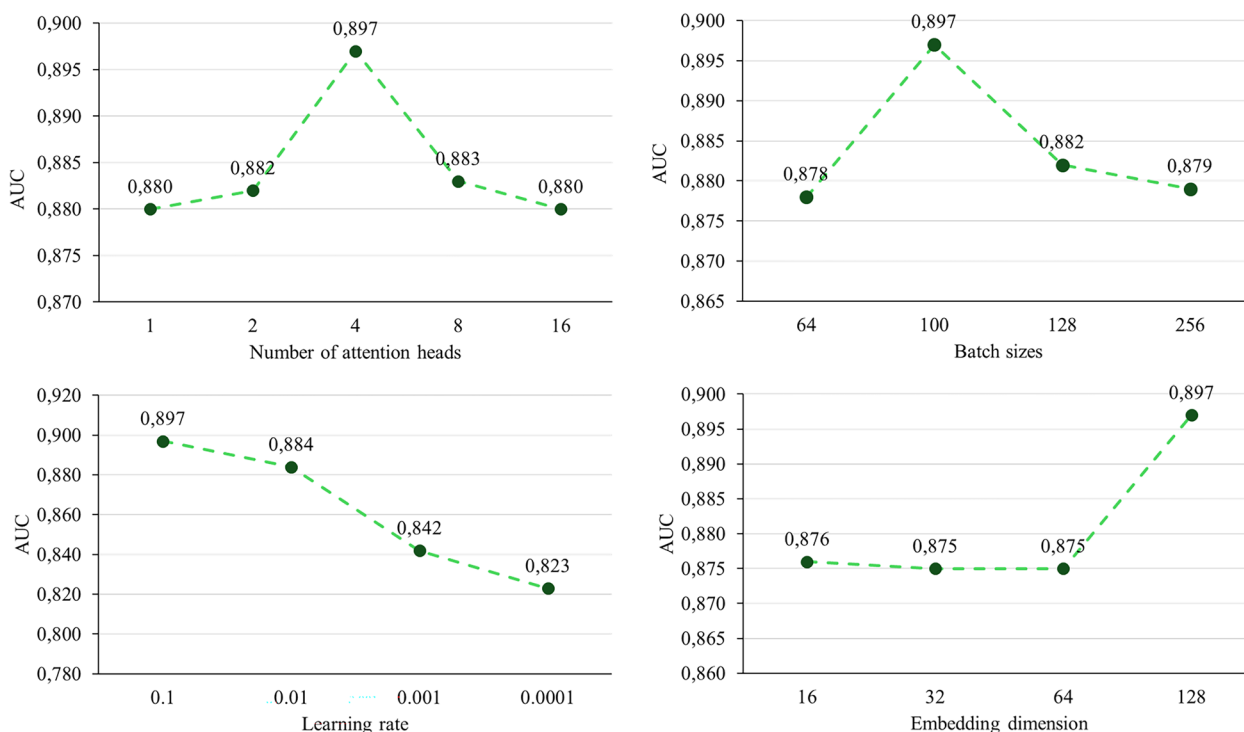


Fig. 6 Experimentations of hyperparameter analysis

$$F1 - Score = 2 \times \frac{PRE \times REC}{PRE + REC} \quad (14)$$

where TP is the positive samples correctly identified, TN is the negative samples correctly identified, FP is the negative samples incorrectly labeled and FN is the positive samples incorrectly labeled.

Hyperparameters and environment setup

We optimize the hyperparameters for GTransCYPs model, as outlined in Table 1, by conducting a search for the best parameters on the training process. The experiment was executed utilizing NVIDIA GTX 1080 Ti×4 and PyTorch. The detailed of the experimental environment is provided in Table 2, and the training was conducted of 20 epochs.

Performance of the proposed model

The performances of GTransCYPs are shown in Table 3. The model demonstrates accurate predictions with BA scores ranging from 0.770 to 0.886; the MCC scores ranging 0.534 to 0.770; REC scores ranging from 0.823 to 0.916; PRE scores ranging from 0.677 to 0.924; and finally, F1-score ranging from 0.743 to 0.920.

Ablation study

We conducted an ablation study to assess the effectiveness of the attention pooling in the GTransCYPs model. The variant models, labeled as model-A, model-B, model-C, and model-D, consist of different pooling techniques: model-A is the GTransCYPs with global mean pooling; Model-B combines global mean pooling and global maximum pooling within the GTransCYPs framework; Model-C utilizes top- k pooling in the GTransCYPs; and in Model-D, the GTransCYPs integrates an attention mechanism into the pooling layers.

From Table 4 it is evident that utilizing attention pooling in the GTransCYPs model outperforms all other pooling schemes in 1A2 dataset in terms of BA, MCC, and F1-score evaluation. We can conclude that the implementation of GTransCYPs with attention pooling is effective due to its superior performance. This can be attributed to the capability of the proposed model to extract the most significant information from each node and edge within the molecular data of CYP450 five isozymes.

Hyperparameter sensitivity

In this study, we investigated the influence of various hyperparameters on the performance of our proposed

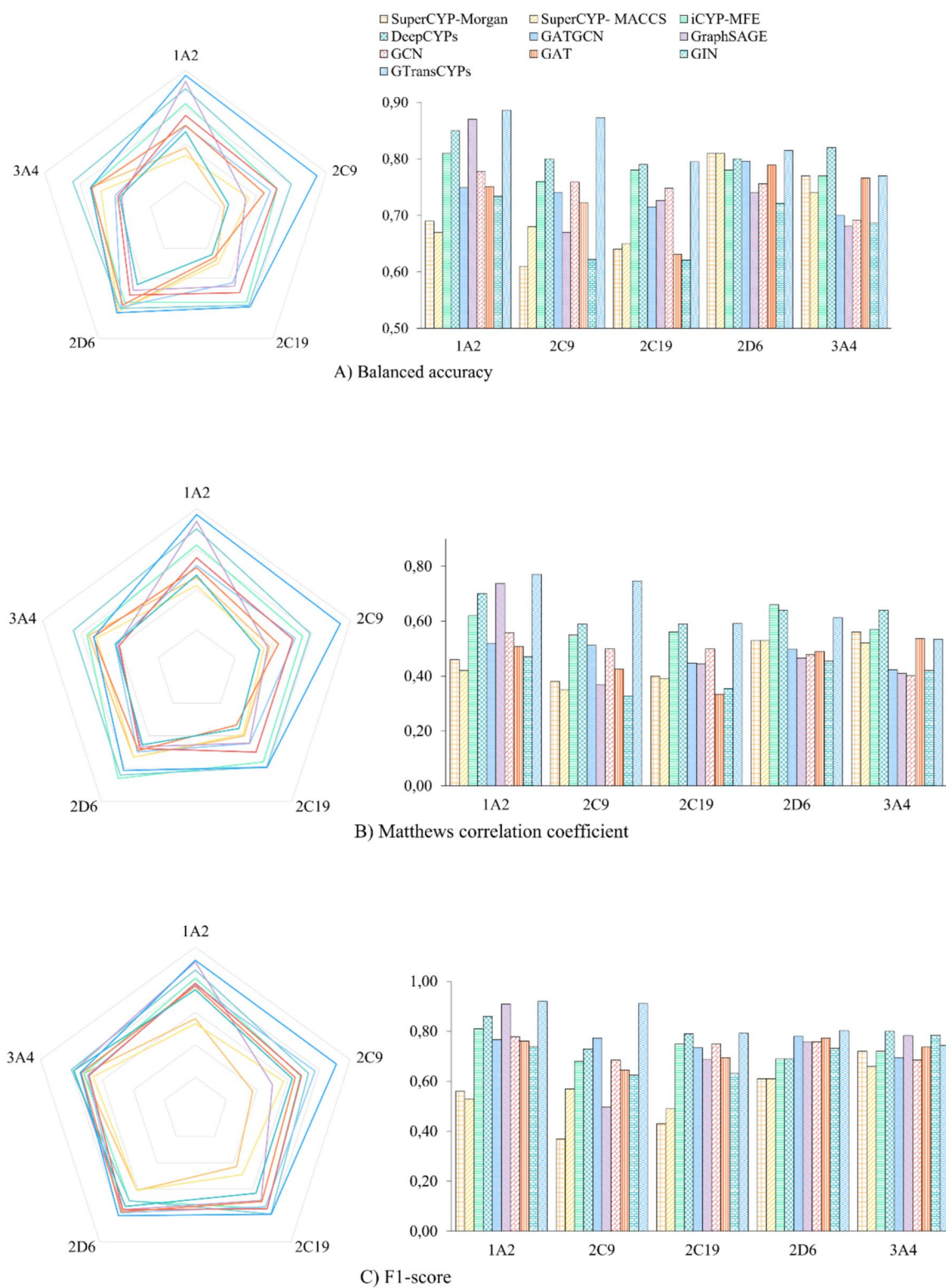


Fig. 7 Comparison of methods across five isozyme CYP450 inhibition datasets using **(A)** BA, **(B)** MCC, and **(C)** F1-Score evaluation

Table 5 The comparison of BA between GTransCYPs and other methods for predicting five CYP450 isozymes inhibition

Models	1A2	2C9	2C19	2D6	3A4
SuperCYP-Morgan [30]	0.690	0.610	0.640	0.810	0.770
SuperCYP-MACCS [30]	0.670	0.680	0.650	0.810	0.740
iCYP-MFE [16]	0.810	0.760	0.780	0.780	0.770
DEEPCYPs [23]	0.850	0.800	0.790	0.800	0.820
GAT_GCN [22]	0.749	0.741	0.715	0.796	0.700
GraphSAGE	0.870	0.670	0.726	0.740	0.681
GCN	0.778	0.759	0.748	0.756	0.692
GAT	0.751	0.723	0.631	0.789	0.766
GIN	0.734	0.622	0.621	0.721	0.687
GTransCYPs	0.886	0.873	0.795	0.815	0.770

Bold denotes the highest value

Table 6 The comparison of MCC between GTransCYPs and other methods for predicting five CYP450 isozymes inhibition

Models	1A2	2C9	2C19	2D6	3A4
SuperCYP-Morgan [30]	0.460	0.380	0.400	0.530	0.560
SuperCYP-MACCS [30]	0.420	0.350	0.390	0.530	0.520
iCYP-MFE [16]	0.620	0.550	0.560	0.660	0.570
DEEPCYPs [23]	0.700	0.590	0.590	0.640	0.640
GAT_GCN [22]	0.519	0.512	0.447	0.498	0.423
GraphSAGE	0.738	0.369	0.445	0.466	0.410
GCN	0.557	0.499	0.499	0.478	0.401
GAT	0.508	0.425	0.334	0.490	0.536
GIN	0.471	0.327	0.355	0.455	0.420
GTransCYPs	0.770	0.746	0.591	0.612	0.534

Bold denotes the highest value

Table 7 The F1-Score comparison between GTransCYPs and other methods for predicting five CYP450 isozymes inhibition

Models	1A2	2C9	2C19	2D6	3A4
SuperCYP-Morgan [30]	0.560	0.370	0.430	0.610	0.720
SuperCYP-MACCS [30]	0.530	0.570	0.490	0.610	0.660
iCYP-MFE [16]	0.810	0.680	0.750	0.690	0.720
DEEPCYPs [23]	0.860	0.730	0.790	0.690	0.800
GAT_GCN [22]	0.767	0.773	0.734	0.780	0.695
GraphSAGE	0.909	0.498	0.687	0.757	0.783
GCN	0.778	0.685	0.750	0.758	0.686
GAT	0.761	0.645	0.695	0.773	0.737
GIN	0.739	0.625	0.632	0.732	0.785
GTransCYPs	0.920	0.911	0.792	0.803	0.743

Bold denotes the highest value

model. The insights derived from our analysis of key parameters' sensitivity, including number of attention heads, batch size, learning rate, and representation embedding dimensions throughout the training process, are illustrated in Fig. 6.

Effect of the number of attention heads

We explored the effect of the number of attention heads on GTransCYPs performance. The experimental results indicate that the area under the ROC curve (AUC) score increased from 0.880 with 1 attention head to its peak at 0.897 with 4 attention heads. However, subsequent to reaching this peak, there was a decline in the AUC score to 0.883 with 8 attention heads, and further dropped to 0.880 with 16 attention heads. The decrease in performance with larger numbers of attention heads may be attributed to excessive model complexity and information redundancy. Therefore, the optimal number of attention heads is 4, as it provides a balanced trade-off between relevant information and complexity.

Effect of the batch size number

We investigated the impact of batch size on model performance, where the highest AUC score was observed with a batch size of 100, reaching 0.897, whereas batch sizes of 64 and 128 yielded slightly lower AUC scores of 0.878 and 0.882, respectively. However, a more substantial decrease in AUC score was noted with a batch size of 256, dropping to 0.879. This decline in performance with a batch size of 256 can be attributed to the reduction in sample size per iteration, potentially disrupting model convergence and resulting in less accurate information extraction from the data. Thus, in this study, a batch size of 100 emerges as the optimal choice for enhancing model performance.

Effect of the learning rate

The analysis indicates that the learning rate affects the AUC score, reaching its highest peak at 0.1 with a value of 0.897. However, there is a decrease in model performance, with the AUC score dropping to 0.823 at a learning rate of 0.0001. Reducing the learning rate can diminish the model's ability to identify important patterns in the data, underscoring the importance of selecting the appropriate learning rate to optimize model performance.

Effect of the size of embedding dimension

We can observe that the AUC score tends to increase with the increase in the embedding dimension size, reaching its peak at a dimension size of 128 with a value

Table 8 Representative molecular structures of inhibitors and non-inhibitors for each CYP isozyme

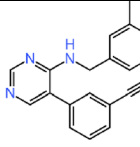
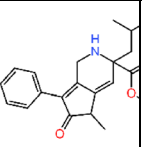
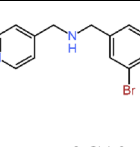
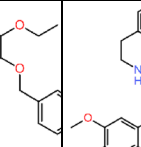
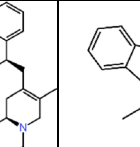
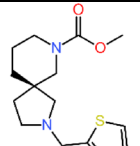
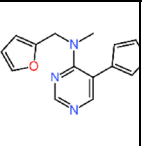
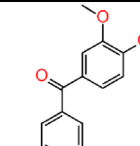
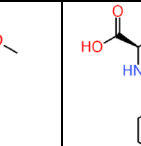
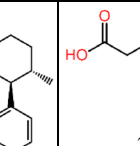
Inhibitors	 1A2 SID: 4239706	 2C9 SID: 17385329	 2C19 SID: 17413919	 2D6 SID: 4253788	 3A4 SID: 11111147
Non-inhibitors	 1A2 SID: 4238458	 2C9 SID: 4240351	 2C19 SID: 4250180	 2D6 SID: 26751174	 3A4 SID: 4253087

Table 9 Comparison of GTransCYPs prediction performance with the other models

SID	IUPAC	Prediction confidence inhibition (%)				
		GIN	GAT	GCN	GraphSAGE	GTransCYPs
4239706	3-[4-[(3-methylphenyl)methylamino]pyrimidin-5-yl]benzoxazole	54.57	60.58	85.39	75.72	90.35
17385329	methyl 5-methyl-3-(2-methylpropyl)-6-oxo-7-phenyl-2,5-dihydro-1H-cyclopenta[c]pyridine-3-carboxylate	59.64	71.81	79.86	83.6	89.27
17413919	N-[[3-bromo-5-ethoxy-4-[(4-fluorophenyl)methoxy]phenyl]methyl]-1-pyridin-4-ylmethanamine;hydrochloride	68.70	55.19	61.97	51.86	69.41
4253788	(11bR)-3-ethyl-9,10-dimethoxy-2-[[[(1R)-1,2,3,4-tetrahydroisoquinolin-1-yl]methyl]-4,6,7,11b-tetrahydro-1H-benzo[a]quinolizine	75.83	57.52	54.34	64.90	79.07
11111147	tert-butyl 3-[4-(1,3-benzodioxol-5-yl)-4-hydroxybutyl]pyrrole-1-carboxylate	44.92	66.49	59.28	51.88	55.78
4238458	methyl (5S)-2-(1,3-thiazol-2-ylmethyl)-2,7-diazaspiro[4.5]decane-7-carboxylate	40.05	46.05	32.53	16.52	08.92
4240351	5-(furan-3-yl)-N-(furan-2-ylmethyl)-N-methylpyrimidin-4-amine	41.54	27.18	40.46	21.83	15.99
4250180	bis(3,4-dimethoxyphenyl)methanone	36.49	19.99	19.70	45.18	17.88
26751174	(2R,5S,6S)-6-(4-ethenylphenyl)-5-methylpiperidine-2-carboxylic acid	44.18	37.15	38.54	29.86	12.08
4253087	2-(2-amino-9-propylpurin-6-yl)sulfanylacetic acid	51.65	49.02	23.24	27.51	47.25

of 0.897. This indicates an enhancement in the model's ability to capture crucial information from the data as the embedding dimension size grows. However, it is important to note that increasing the embedding dimension size will result in an increase in the number of parameters in the model, consequently prolonging the training time and requiring more computational resources.

Comparison of GTransCYPs with existing methods

To ensure a comprehensive assessment of our proposed model, we compared GTransCYPs with existing methods such as SuperCYP [30], iCYP-MFE [16], DeepCYPs

[23], GAT_GCIN [22], GraphSAGE [31], GAT [32], GCN [33], and GIN [34]. Figure 7 shows a comparison between GTransCYPs and other methods in predicting CYP450 inhibition across five isozyme datasets. We can see that GTransCYPs outperforms advanced methods on most datasets. Table 5 reports that the GraphSAGE model outperforms the GAT_GCIN, DeepCYPs, GAT, GCN, and GIN models with a score of 0.870 on dataset 1A2 in terms of BA values. However, a notable performance improvement is observed in the GTransCYPs model, which surpasses GraphSAGE with an increase of 1.8%. For datasets 2C9, 2C19, and 2D6, GTransCYPs

Step 1: select menu "Server"

Step 2: select CYP450 type

Step 3: Upload a CSV file containing the list molecules for virtual screening

Step 4: click "Prediction"

Results

CYP450 1A2 Inhibition prediction output

	Molecule SMILES Sequence	PConfinh (%)	Inhibitor
0	<chem>CSc1nc2nc3c(c(=O)n2)[nH]1)CN(Cc1cccc1)CC3</chem>	56.96	✓
1	<chem>O=C(c1cnccn1)N1CCC2(CC1)CN(c1ccccn1)C2</chem>	44.96	✗
2	<chem>Cc1ccc(NC(=S)NC2CCCC2)cc1</chem>	66.72	✓
3	<chem>C=CCNCCOCCOc1ccc(C)cc1[N+](=O)[O-]</chem>	90.35	✓
4	<chem>NCCSC[C@H](N)C(=O)O</chem>	2.75	✗
5	<chem>C=CCn1c(SCc2ccc(C#N)cc2)nnc1-c1ccc(C)cc1</chem>	91.69	✓
6	<chem>C#CCCCO/N=C1/C[C@@H](O)[C@@H](O)[C@H]2[C@@H]1CC[C@@H]2</chem>	2.72	✗
7	<chem>Cc1noc(C)c1C(=O)N1CCC2(CC1)CN(Cc1cc(C(F)(F)F)cc(C(F)(F)F)c1)C2</chem>	0.94	✗
8	<chem>O=C(O)C1C2C=CC3(CN(Cc4cccn4)C(=O)C13)O2</chem>	3.39	✗

PConfinh (%): GTransCYPs model's Prediction confidence for molecule to be an inhibitor.

Fig. 8 The interface of the webserver

outperformed all models with performance improvements of 9%, 0.6%, and 0.6% respectively. Compared to other methods, GTransCYPs achieved the highest MCC scores with improvements of 4.3%, 26.4%, and 0.2% for datasets 1A2, 2C9, and 2C19 respectively, as detailed in Table 6. The proposed model also demonstrates good performance compared to other models for F1-score evaluation. Table 7 presents that GTransCYPs model outperforms all other methods across most datasets, with an improvement of 1.2%, 17.9%, 0.3% and 2.9% for datasets 1A2, 2C9, 2C19, and 2D6, respectively. Table 8 presents representative molecular structures of both inhibitors and non-inhibitors for the CYP450 isoenzymes. Meanwhile, Table 9 displays the predicted inhibition activity of various compounds against CYP450 isoenzymes using four different methods: GIN, GAT, GCN, GraphSAGE, and GTransCYPs. Although all models predicted

correctly, the proposed model demonstrates a high prediction confidence for inhibition. We can see that the GTransCYPs achieved scores of 90.35 and 89.27 when predicting inhibitory activities of SDI 4239706 and SDI 17385329, respectively. These results indicate a significant difference compared to other models.

Webserver

GTransCYPs source code is available at <https://github.com/zonwoo/GTransCYPs> and can be locally hosted using the streamlit platform (<https://www.streamlit.io>). This platform has been designed to provide support for researchers and practitioners in the fields of chemistry, biology, and pharmacology in their drug development research, especially concerning online prediction of CYP450 inhibition. The interface is designed to ensure simplicity and user-friendliness, thereby enhancing the

experience for both novice and experienced users. Figure 8 presents an overview of the user interface along with examples of processing snippets.

Conclusion

Predicting CYP450 inhibition is one of the key challenges in drug research and holds significant implications across various clinical applications. This study introduced a novel graph representation learning model GTransCYPs for CYP450 inhibition prediction. GTransCYPs first learns low-dimensional molecular representations and constructs topological graphs by integrating attention mechanisms and transformer feature architectures. Additionally, it integrates graph pooling to simplify the complexity of graph structures by preserving a designated number of informative nodes within each subgraph. This approach amplifies efficiency and accentuates the focus on pertinent information essential for predicting CYP450 inhibitors. According to the experimental results, GTransCYPs achieves competitive performance compared to existing methods. Ablation experiments provide additional clarity on the key roles of pooling layers in boosting the predictive capabilities of the proposed method. However, there are still a few improvements that should be considered. Due to the limited size of the publicly available dataset, particularly in the case of the 3A4 and 2D6 datasets, there is a significant disparity between the positive and negative classes. This imbalance heavily impacts the dataset, resulting in suboptimal model performance. Additionally, model interpretability remains a limitation, as it is crucial to understand how the proposed model learns patterns in molecular substructures and identifies molecules with potential inhibitory activity on various enzymes during the molecule design process. In future work, we will explore various strategies, such as transfer learning and semi-supervised learning approaches based on graph representations. Furthermore, we will focus on model interpretability to provide deeper insights into how the model learns patterns in molecular substructures.

Acknowledgements

Not applicable.

Author contributions

Conceptualization: C.Z., J.D.K.; Methodology: C.Z.; Investigation: C.Z., S.N.N., S.M.; Formal Analysis: C.Z., J.D.K., S.N.N., S.M.; Visualization: C.Z., S.N.N., J.D.K., S.M.; Supervision: J.D.K.; Writing—original draft: C.Z., S.N.N.; Writing—review C.Z., J.D.K., S.N.N., S.M.

Funding

This research was supported by the MSIT (Ministry of Science, ICT), Korea, under the National Program for Excellence in SW supervised by the IITP (Institute of Information & communications Technology Planning & Evaluation) in 2024 (2024-0-00023).

Availability of data and materials

The CYP450 datasets of this work can be found at <https://tdcommons.ai> (therapeutics data commons). The source codes are available at <https://github.com/zonwoo/GTransCYPs>.

Declarations

Competing interests

The authors declare no competing interests.

Received: 7 May 2024 Accepted: 10 October 2024

Published online: 29 October 2024

References

1. Rendic SP, Guengerich FP (2021) Human Family 1–4 cytochrome P450 enzymes involved in the metabolic activation of xenobiotic and physiological chemicals: an update. *Arch Toxicol* 95:395–472
2. Peter GF (1994) Catalytic selectivity of human cytochrome P450 enzymes: relevance to drug metabolism and toxicity. *Toxicol Lett* 70:133–138
3. Zhao M, Ma J, Li M et al (2021) Cytochrome P450 enzymes and drug metabolism in humans. *IJMS* 22:12808
4. Song Y, Li C, Liu G et al (2021) Drug-metabolizing cytochrome P450 enzymes have multifarious influences on treatment outcomes. *Clin Pharmacokinet* 60:585–601
5. Yu M-S, Lee H-M, Park A et al (2018) In silico prediction of potential chemical reactions mediated by human enzymes. *BMC Bioinform* 19:207
6. Evans WE, Relling MV (1999) Pharmacogenomics: translating functional genomics into rational therapeutics. *Science* 286:487–491
7. Di L (2014) The role of drug metabolizing enzymes in clearance. *Expert Opin Drug Metab Toxicol* 10:379–393
8. Wang Y, Bryant SH, Cheng T et al (2017) PubChem BioAssay: 2017 update. *Nucleic Acids Res* 45:D955–D963
9. Lee JH, Basith S, Cui M et al (2017) In silico prediction of multiple-category classification model for cytochrome P450 inhibitors and non-inhibitors using machine-learning method. *SAR QSAR Environ Res* 28:863–874
10. Kato H (2020) Computational prediction of cytochrome P450 inhibition and induction. *Drug Metab Pharmacokinet* 35:30–44
11. Plonka W, Stork C, Sicho M et al (2021) CYPlebrity: machine learning models for the prediction of inhibitors of cytochrome P450 enzymes. *Bioorg Med Chem* 46:116388
12. Xu M, Lu Z, Wu Z et al (2023) Development of In silico models for predicting potential time-dependent inhibitors of cytochrome P450 3A4. *Mol Pharmaceutics* 20:194–205
13. Wu Z, Lei T, Shen C et al (2019) ADMET evaluation in drug discovery. 19. Reliable prediction of human cytochrome P450 inhibition using artificial intelligence approaches. *J Chem Inf Model* 59:4587–4601
14. Goldwasser E, Laurent C, Lagarde N et al (2022) Machine learning-driven identification of drugs inhibiting cytochrome P450 2C9. *PLoS Comput Biol* 18:e1009820
15. Wang N-N, Wang X-G, Xiong G-L et al (2022) Machine learning to predict metabolic drug interactions related to cytochrome P450 isozymes. *J Cheminform* 14:23
16. Nguyen-Vo T-H, Trinh QH, Nguyen L et al (2022) iCYP-MFE: identifying human cytochrome P450 inhibitors using multitask learning and molecular fingerprint-embedded encoding. *J Chem Inf Model* 62:5059–5068
17. Su B-H, Tu Y, Lin C et al (2015) Rule-based prediction models of cytochrome P450 inhibition. *J Chem Inf Model* 55:1426–1434
18. Tang M, Li B, Chen H (2023) Application of message passing neural networks for molecular property prediction. *Curr Opin Struct Biol* 81:102616
19. Buterez D, Janet JP, Kiddle SJ et al (2024) Transfer learning with graph neural networks for improved molecular property prediction in the multi-fidelity setting. *Nat Commun* 15:1517
20. Wei Z, Zhao C, Zhang M et al (2024) Meta-DHGNN: method for CRS-related cytokines analysis in CAR-T therapy based on meta-learning

- directed heterogeneous graph neural network. *Brief Bioinform* 25:bbae104
21. Meller A, Ward M, Borowsky J et al (2023) Predicting locations of cryptic pockets from single protein structures using the PocketMiner graph neural network. *Nat Commun* 14:1177
 22. Qiu M, Liang X, Deng S et al (2022) A unified GCNN model for predicting CYP450 inhibitors by using graph convolutional neural networks with attention mechanism. *Comput Biol Med* 150:106177
 23. Ai D, Cai H, Wei J et al (2023) DEEPCYPs: a deep learning platform for enhanced cytochrome P450 activity prediction. *Front Pharmacol* 14:1099093
 24. Gillioz A, Riesen K (2023) Graph-based pattern recognition on spectral reduced graphs. *Pattern Recognition* 144:109859
 25. Chen B, Barzilay R, Jaakkola T. Path-augmented graph transformer network. 2019.
 26. Ramsundar B, Eastman P, Walters P et al (2019) Deep learning for the life sciences. O'Reilly Media, Sebastopol
 27. RDKit: Open-source cheminformatics. <https://zenodo.org/record/3732262>.
 28. Veith H, Southall N, Huang R et al (2009) Comprehensive characterization of cytochrome P450 isozyme selectivity across chemical libraries. *Nat Biotechnol* 27:1050–1055
 29. Huang K, Fu T, Gao W et al (2022) Artificial intelligence foundation for therapeutic science. *Nat Chem Biol* 18:1033–1036
 30. Banerjee P, Dunkel M, Kemmler E et al (2020) SuperCYPsPred—a web server for the prediction of cytochrome activity. *Nucleic Acids Res* 48:W580–W585
 31. Guo Z, Yu W, Zhang C et al. GraSeq: graph and sequence fusion learning for molecular property prediction. In: *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. Virtual Event Ireland: ACM, 2020, 435–43.
 32. Feng Y-Y, Yu H, Feng Y-H et al (2022) Directed graph attention networks for predicting asymmetric drug–drug interactions. *Brief Bioinform* 23:bbac151
 33. Zhao T, Hu Y, Valsdottir LR et al (2021) Identifying drug–target interactions based on graph convolutional network and deep neural network. *Brief Bioinform* 22:2141–2150
 34. Zheng K, Zhao H, Zhao Q et al (2022) NASMDR: a framework for miRNA–drug resistance prediction using efficient neural architecture search and graph isomorphism networks. *Brief Bioinform* 23:bbac338

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.