D-CyPre: A machine learning-based tool for accurate prediction of site of metabolism by human CYP450 enzyme (#92258)

First submission

Guidance from your Editor

Please submit by **12 Dec 2023** for the benefit of the authors (and your token reward) .

Structure and Criteria
Please read the 'Structure and Criteria' page for general guidance.

Raw data check
Review the raw data.

Image check
Check that figures and images have not been inappropriately manipulated.

If this article is published your review will be made public. You can choose whether to sign your review. If uploading a PDF please remove any identifiable information (if you want to remain anonymous).

Files

Download and review all files from the materials page.

4 Figure file(s)
4 Table file(s)
1 Raw data file(s)
1 Other file(s)

Structure and Criteria

Structure your review

The review form is divided into 5 sections. Please consider these when composing your review:

1. **BASIC REPORTING**
2. **EXPERIMENTAL DESIGN**
3. **VALIDITY OF THE FINDINGS**
4. General comments
5. Confidential notes to the editor

You can also annotate this PDF and upload it as part of your review

When ready submit online.

Editorial Criteria

Use these criteria points to structure your review. The full detailed editorial criteria is on your guidance page.

BASIC REPORTING

- Clear, unambiguous, professional English language used throughout.
- Intro & background to show context. Literature well referenced & relevant.
- Structure conforms to PeerJ standards, discipline norm, or improved for clarity.
- Figures are relevant, high quality, well labelled & described.
- Raw data supplied (see PeerJ policy).

EXPERIMENTAL DESIGN

- Original primary research within Scope of the journal.
- Research question well defined, relevant & meaningful. It is stated how the research fills an identified knowledge gap.
- Rigorous investigation performed to a high technical & ethical standard.
- Methods described with sufficient detail & information to replicate.

VALIDITY OF THE FINDINGS

- Impact and novelty not assessed. *Meaningful* replication encouraged where rationale & benefit to literature is clearly stated.
- All underlying data have been provided; they are robust, statistically sound, & controlled.
- Conclusions are well stated, linked to original research question & limited to supporting results.

Standout reviewing tips

The best reviewers use these techniques

Tip	*Example*
Support criticisms with evidence from the text or from other sources	*Smith et al (J of Methodology, 2005, V3, pp 123) have shown that the analysis you use in Lines 241-250 is not the most appropriate for this situation. Please explain why you used this method.*
Give specific suggestions on how to improve the manuscript	*Your introduction needs more detail. I suggest that you improve the description at lines 57- 86 to provide more justification for your study (specifically, you should expand upon the knowledge gap being filled).*
Comment on language and grammar issues	*The English language should be improved to ensure that an international audience can clearly understand your text. Some examples where the language could be improved include lines 23, 77, 121, 128 – the current phrasing makes comprehension difficult. I suggest you have a colleague who is proficient in English and familiar with the subject matter review your manuscript, or contact a professional editing service.*
Organize by importance of the issues, and number your points	*1. Your most important issue* *2. The next most important item* *3. …* *4. The least important points*
Please provide constructive criticism, and avoid personal opinions	*I thank you for providing the raw data, however your supplemental files need more descriptive metadata identifiers to be useful to future readers. Although your results are compelling, the data analysis should be improved in the following ways: AA, BB, CC*
Comment on strengths (as well as weaknesses) of the manuscript	*I commend the authors for their extensive data set, compiled over many years of detailed fieldwork. In addition, the manuscript is clearly written in professional, unambiguous language. If there is a weakness, it is in the statistical analysis (as I have noted above) which should be improved upon before Acceptance.*

D-CyPre: A machine learning-based tool for accurate prediction of site of metabolism by human CYP450 enzyme

Haolan Yang [1, 2], **Jie Liu** [2], **Kui Chen** [1, 2], **Shiyu Cong** [1, 2], **Shengnan Cai** [1, 2], **Yueting Li** [1, 2], **Zhixin Jia** [2], **Hao Wu** [1, 2], **Tianyu Lou** [1, 2], **Zuying Wei** [1, 2], **Xiaoqin Yang** [1, 2], **Hongbin Xiao** [Corresp. 2]

[1] School of Chinese Materia Medica, Beijing University of Chinese Medicine, Beijing, China

[2] Beijing University of Chinese Medicine, Research Center of Chinese Medicine Analysis and Transformation, Beijing, China

Corresponding Author: Hongbin Xiao
Email address: hbxiao69@163.com

The advancement of graph neural networks (GNNs) has enhanced the accuracy of predicting metabolic sites. However, research in this domain remains scarce, with only a few preliminary investigations conducted thus far on the efficacy of fundamental GNNs. Moreover, research indicates that the fusion of GNNs with XGBOOST exhibits superior performance, yet such experimentation has not been attempted in the realm of metabolic site prediction. Additionally, most metabolic site prediction tasks only focus on bonds and atoms, often neglecting information on the overall molecular structure. Even GNNs merely depict the local environment of atoms. Therefore, it is imperative to establish a more rational and efficient model for predicting metabolic sites. In this study, we have devised a novel tool named D-CyPre, which amalgamates atom, bond, and molecule information via two directed message-passing neural networks (D-MPNN) and employs XGBOOST to predict the metabolic sites (SOM) of nine cytochrome P450 (CYP450) enzymes. D-CyPre has two modes: Precision Mode, which emphasizes high precision, and Recall Mode, which emphasizes high recall, catering to different user needs. In both the validation and test sets, D-CyPre's performance consistently surpasses that of existing models. Our results indicate that the features of molecules may play a positively impactful role in predicting metabolic sites.

D-CyPre: A machine learning-based tool for accurate prediction of site of metabolism by human CYP450 enzyme

Haolan Yang[1,2], Jie Liu[2], Kui Chen[1,2], Shiyu Cong[1,2], Shengnan Cai, Yueting Li[1,2], Zhixin Jia[2], Hao Wu[1,2], Tianyu Lou[1,2], Zuying Wei[1,2], Xiaoqin Yang[1,2], Hongbin Xiao[2]*

[1] School of Chinese Materia Medica, Beijing University of Chinese Medicine, Beijing, China
[2] Research Center of Chinese Medicine Analysis and Transformation, Beijing University of Chinese Medicine, Beijing, China

Corresponding Author:
Hongbin Xiao[1,2]
Intersection of Yangguang South Street and Baiyang East Road, Beijing, Beijing, 102488, China
Email address: hbxiao69@163.com

ABSTRACT

The advancement of graph neural networks (GNNs) has enhanced the accuracy of predicting metabolic sites. However, research in this domain remains scarce, with only a few preliminary investigations conducted thus far on the efficacy of fundamental GNNs. Moreover, research indicates that the fusion of GNNs with XGBOOST exhibits superior performance, yet such experimentation has not been attempted in the realm of metabolic site prediction. Additionally, most metabolic site prediction tasks only focus on bonds and atoms, often neglecting information on the overall molecular structure. Even GNNs merely depict the local environment of atoms. Therefore, it is imperative to establish a more rational and efficient model for predicting metabolic sites. In this study, we have devised a novel tool named D-CyPre, which amalgamates atom, bond, and molecule information via two directed message-passing neural networks (D-MPNN) and employs XGBOOST to predict the metabolic sites (SOM) of nine cytochrome P450 (CYP450) enzymes. D-CyPre has two modes: Precision Mode, which emphasizes high precision, and Recall Mode, which emphasizes high recall, catering to different user needs. In both the validation and test sets, D-CyPre's performance consistently surpasses that of existing models. Our results indicate that the features of molecules may play a positively impactful role in predicting metabolic sites.

INTRODUCTION

Cytochrome P450 (CYP450) enzymes are responsible for the metabolism of approximately 90% of FDA-approved medicines and play a vital role in the Phase I metabolism of drugs(Nebert & Russell, 2002). As the primary and most convenient route of administration, oral intake

invariably results in alterations to the molecular structures of drugs(Hou et al., 2007; Xu et al., 2012; Wang & Hou, 2015). The metabolism of drugs is closely linked to their bioavailability, bioactivity, and toxicology. When a drug is rapidly metabolized upon entering the body, only a small amount of the original compound remains, leading to reduced bioactivity and bioavailability. Furthermore, if the metabolites produced are toxic, drug use will be restricted. Hence, predicting how CYP450 will metabolize drugs can help us modify the drug's molecular structure to avoid such undesired situations. In conclusion, predicting drug metabolism by CYP450 isoforms is crucial for drug design and discovery(Jianing et al., 2011).

Several *in silico* metabolism prediction tools have been developed to discover and design drugs more effective, such as CyProduct(Tian et al., 2021), CypReact(Tian et al., 2018), FAME2(Šícho et al., 2017) and FAME3(Šícho et al., 2019a). However, all these models rely on fixed rules to generate the features of the site of metabolism (SOM) or bond of metabolism (BOM). While graph neural networks (GNNs) are less prevalent *in silicon* metabolism prediction tasks, they have already demonstrated their efficacy in replacing conventionally handcrafted molecular features generated by fixed rules in other molecular-related research domains. Recently, GNNs have shown a promising effect on molecular property prediction(Gilmer et al., 2017; Yang et al., 2019) and drug discovery(Stokes et al., 2020; Jin et al., 2021). The commonly used GNNs in these studies are message passing neural networks (MPNN) (Gilmer et al., 2017; Jo et al., 2020) and directed MPNN (D-MPNN)(Yang et al., 2019; Stokes et al., 2020; Jin et al., 2021; Han et al., 2022). Both networks use message-passing to aggregate the chemical information from the entire molecule and learn how to generate better features. The difference between them lies in the types of messages: MPNN aggregates information from related vertices (atoms), while D-MPNN aggregates information from directed edges (bonds). Compared to the MPNN, the D-MPNN can avoid loops in message-passing (Yang et al., 2019).

In many studies predicting SOMs or BOMs, models often include information about neighboring atoms or bonds when creating features for atoms or bonds (He et al., 2016; Šícho et al., 2017, 2019b; de Bruyn Kops et al., 2019, 2021; Tian et al., 2021). However, this step is very subjective, and it is difficult to determine which features of adjacent structures are required by the model. So there is room for improvement in models that are based on these features. In contrast, the D-MPNN requires only the features of the target atom or bond, and which features of neighboring structures are important will be determined by the neural network. Also, the neural network does not just screen the features but transforms the features, which may generate some new features that are more effective for determining SOMs. In summary, the D-MPNN has shown excellent results in other fields and has an objective and powerful ability for feature generation. We believe that it may achieve better results than existing models *in silico* metabolism prediction.

We have taken note of recent studies wherein researchers have systematically examined the performance of GNNs in predicting metabolic sites (Porokhin, Liu & Hassoun, 2023). However, the GNNs that was scrutinized lacks the incorporation of the novel D-MPNN and has not evolved into a user-friendly tool for scientific researchers. Furthermore, training stable models

for molecular property prediction using a multi-layer perceptron may prove to be challenging. Study have suggested that employing a GNNs in conjunction with XGBOOST for training yields superior predictive performance (Deng et al., 2021). Furthermore, the overall structure of the molecule is a crucial factor. This study also examines the impact of fusing traditional molecular features or features generated based on D-MPNN with those generated from the bonds and atoms within the molecule using D-MPNN. This study holds distinctive significance in terms of developing a novel metabolic site prediction model with better performance or aiding non-computational personnel in their research within the field of metabolism.

In this study, we established D-CyPre, an *in silico* metabolism predictor capable of predicting any of the nine most significant human CYP450 enzymes (Phase I metabolism) (Zanger & Schwab, 2013). As shown in Figure 1, D-CyPre can be divided into two parts. The first part is to generate the features by D-MPNN, and the second part is to predict metabolic sites by these features. Finally, D-CyPre visually displays the predicted results (Figure 1). The darker the red in the figure, the higher the probability of metabolism of this site. Additionally, the probability value is written on the target atom or bonding atom of the target bond. It's worth noting that D-CyPre only displays valuable sites with a probability greater than 50%.

MATERIALS AND METHODS

2.1 Data Sets.

The data set used for training model in this study was EBoMD data set from CyProduct(Tian et al., 2021). This public data set includes BOMs of 679 substrates on nine of the most important human CYP450 isoforms (CYP1A2, CYP2A6, CYP2B6, CYP2C8, CYP2C9, CYP2C19, CYP2D6, CYP2E1, CYP3A4) created from the Zaretzki Data set (Zanger & Schwab, 2013; Zaretzki, Matlock & Swamidass, 2013). The Zaretzki Data set has been used in several related studies of *in silico* metabolism predictor (Tian et al., 2018, 2021; Šícho et al., 2019a; Dang et al., 2020). Tian, S et al. converts SOMs in Zaretzki Data set to BOMs during the creation of the EBoMD, while correcting some errors (Tian et al., 2021). Finally, the EBoMD mainly consists of the following nine Phase I reactions: Oxidation, Cleavage, EpOxidation, Reduction, Hydroxylation, S(sulfur)-Oxidation, N(nitrogen)-Oxidation, P (phosphorus)-Oxidation, and Cyclization (Tian et al., 2021).

To evaluate D-CyPre's performance and compare it with CyProduct's performance we used EBoMD2, which comes from CyProduct and contains 68 extracted reactants and 30 known non-CYP450 reactants as a test data set (Tian et al., 2021).

2.2 Atoms and Bonds of Metabolism

CyProduct came up with BOM, and Tian, S et al. argue BOM is more clearly defined and classified more systematically than SOM (Tian et al., 2021). According to the structure of D-CyPre, a new definition is made based on the BOM. This definition does not necessarily perform well in other models, but it is suitable for D-CyPre. Because with regards to D-Cypre, the features of atoms and bonds are both descended to the same dimensions by neural networks, there may be some common knowledge about the features of both. In this study, we still refer to these defined atoms and bonds as SOMs. The specific rules are described as follows:

(1) i-j: i and j represent any two non-H atoms currently connected by an existing chemical bond. We define the bond formed by these two atoms as the SOM that D-CyPre should recognize.

(2) i-H: i represents any non-H atom, and hydrogen atoms on i will be replaced with heteroatoms. We define atom i and the bond formed between i and H as SOMs that D-CyPre should recognize because this reaction involves both i and its bonds with H.

(3) SPN: When new bonds are generated on S, P, or N by sharing their lone pair electrons, we define these atoms as SOMs that D-CyPre needs to recognize because this reaction only involves atoms (S, P, or N).

Instead of creating a model for each type of bond, as CyProduct does (Tian et al., 2021), we used only one model to identify all types of SOMs of one CYP450 isoform. We do not even treat atoms and bonds separately but use the same discriminator to determine whether they are SOMs. The reason why we determine atoms and bonds by the same model is that the information of them can be well crossed and fused in the process of message-pass of D-MPNN, and the model is likely to learn more positive information without distinguishing them. The distribution of SOMs for nine CYP450 isoforms is shown in Table 1.

2.3 Feature Generation.

D-CyPre includes nine atom descriptors and four bond descriptors (Table 2), with details of these descriptors available in Table S1. It is important to note that the data used for training and testing primarily consists of C, H, O, N, S, and P. To prevent a large number of dimensions that cannot be learned, we assign the same value to all other types of atoms when calculating the Atomic Number.

2.4 D-CyPre

D-CyPre consists of D-MPNN and XGBOOST, where D-MPNN outputs the features of atoms and bonds while XGBOOST identifies SOMs based on these features. We will discuss these two structures in detail next.

2.4.1 D-MPNN

The D-MPNN is built based on ComboNet's MPN, which originally came from the Chemprop Software (Yang et al., 2019; Jin et al., 2021) that is open source and available at https://github.com/chemprop/chemprop. First, we're going to fuse the information about the directed bonds and their starting atoms.

$$h_{vw}^{0} = \tau\left(W_a cat\left(x_v, e_{vw}\right)\right) \quad (1)$$

Where $W_a \in \mathbb{R}^{h \times h_a}$ is a learned matrix, $cat\left(x_v, e_{vw}\right) \in \mathbb{R}^{h_a}$ splice together e_{vw}, the feature of a directed bond, and x_v, the feature of the initial atom of the bond. Then, τ is the LeakyReLU activation function (Xu et al., 2015). After that, the message-pass begins

$$m_{vw}^{t+1} = \sum_{k \in \{N(v) \setminus w\}} h_{kv}^{t} \quad (2)$$

$$h_{vw}^{t+1} = drop(\tau\left(h_{vw}^{0} + W_m m_{vw}^{t+1}\right)) \quad (3)$$

Where $W_m \in \mathbb{R}^{h \times h}$ is a learned matrix, and $drop$ is the Dropout layer (Srivastava et al.). The message-pass will be repeated n times, which represents the depth of message-pass, that is, the greater the n, the farther the message will pass. Then, calculate the features of bonds and atoms from the message.

$$F_{vw} = \mathcal{B}(mean(h_{vw}^{n}, h_{wv}^{n})) \quad (4)$$

$$F_v = \mathcal{B}(drop(\tau(W_o cat(x_v, \sum_{W \in N(v)} h_{vw}^{n})))) \quad (5)$$

Where $mean$ is calculate the average value of the two directions of same bond, and $\mathcal{B}$ is the Batch Normalization (Ioffe & Szegedy, 2015). Also, $W_o \in \mathbb{R}^{h \times h_b}$ is a learned matrix and $cat(x_v, \sum_{W \in N(v)} h_{vw}^{n}) \in \mathbb{R}^{h_b}$. Note that the same Batch Normalization layer is used for both atoms and bonds. After that, we feed F_v and F_{vw} into a single-layer neural network, and for each bond and atom, we end up with two values, the positive probability and the negative probability. We then use the cross-entropy to calculate the loss of the model.

$$loss = a \times loss_p + b \times loss_n \quad (6)$$

Where $loss_p$ and $loss_n$ are the loss of atoms and bonds that are truly labeled positive and negative, respectively. Then, a and b are two self-defined parameters, which respectively represent the importance that we attach to the $loss_p$ and $loss_n$. These two parameters are adjusted when training models of the different CYP450 isoforms.

2.4.2 XGBOOST

XGBOOST was proposed by Tianqi Chen (Chen & Guestrin, 2016) and has demonstrated excellent results in several studies (Yu et al., 2019; Chen et al., 2021; Zhang, Hu & Yang, 2022). Daiguo Deng et al. showed that the DMPNN+XGBOOST model can effectively improve the prediction of various molecular properties (Deng et al., 2021). Therefore, this study adopts the similar idea to train models. In general, we trained an XGBOOST model based on F_{vw} and F_v and output Jaccard Score (TP/(TP+FP+FN), Precision (TP/(TP+FP)), Recall (TP/(TP+FN)) and F1 (2 × Precision × Recall/(Precision+Recall)) in each epoch of D-MPNN. The objective and Feval for XGBOOST are set to “binary: logistic” and Jaccard Score, respectively. Other parameters for XGBOOST such as “n estimators”, “reg lambda”, “max depth” and “colsample bytree” are tuned for different isoforms.

2.4.3 Molecular Features

Molecular features play a crucial role in identifying SOMs. For instance, two atoms or bonds in similar conditions may react differently with CYP450 due to their molecular structure, one may react while the other may not. Such bonds or atoms are hard to identify without molecular features. This study considers two types of molecular features. The first type is generated by a new D-MPNN (Yang et al., 2019), while the second type is directly calculated according to specific rules (MolWt; NumHAcceptors; NumHDonors; MolLogP; TPSA; LabuteASA). Details of these descriptors can be found in Table S1.

In this study, when we use the molecular features, the molecular features will be directly concatenated with the features of the atoms and bonds contained in that molecule. Although molecular features are a significant priority, their introduction did not necessarily improve the Jaccard Score of all models in this study. There are two main reasons for this. First, the model used in this study is already complex enough and introducing molecular features may not further improve it or could even cause more severe overfitting. Second, because the data set is not large enough, the model can only learn a small amount of molecular information which may become a disturbance for some isoforms of CYP450. Figure 2 illustrates the structure of D-CyPre that incorporates molecular features.

2.4.4 Precision Mode and Recall Mode

D-CyPre has two modes of high precision and high recall. The difference between the two is that in Precision Mode, XGBOOST's "scale pos weight" is set to the default, while in Recall Mode, this parameter is set to (c × Positive/Negative), where c is a parameter that can be adjusted.

2.4.5 Training model.

For any CYP450 isoform, we divide the EBoMD into a train set and validation set in a ratio of 8:2 (since the features of SOMs are affected by the entire molecular structure, we use molecules rather than SOMs as the minimum unit when dividing the data set). Based on these data, we adjust the model parameters to obtain those with high Jaccard Score in both the training and validation sets.

During this process, we train D-MPNN using data from all isoforms and then train XGBOOST using data from only the target isoform (Figure 3). This improves both Jaccard Score and generalization ability of the model because we believe there is common knowledge among metabolism of nine isoforms. Although learning more knowledge from other isoform may introduce some noise into the model, this knowledge and moderate noise enhance its generalization ability (supplementary files 1). Finally, based on parameters with high Jaccard Score in both training and validation sets, we use the same method to train final D-CyPre and test it with test set.

EXPERIMENTAL RESULTS AND DISCUSSION

3.1 Training model

3.1.1 Precision Mode

Training results are shown in Table S2. The Jaccard Score of D-CyPre-val for nine CYP450 enzymes was higher than that of CyProduct. Similarly, D-CyPre-val showed higher Precision and F1 for eight enzymes other than 2C8. However, since D-CyPre-Val and CyProducts use different validation sets and methods, this result does not prove that D-CyPre necessarily has better predictive power than CyProduct.

3.1.2 Recall Mode

According to results shown in Table S3, D-CyPre-val has higher Jaccard Score, Recall and F1 for nine CYP450 enzymes. This indicates that D-CyPre has good predictive power.

3.2 The results of testing

3.2.1 Precision Mode

The results of our analysis are presented in Table S4. Utilizing Precision Mode, D-CyPre enhances Precision (WAvg) by 39% on the test set in comparison to CyProduct. Likewise, D-CyPre sustains higher Jaccard Score (WAvg) and F1 (WAvg), with increases of 15% and 11%, respectively. The outcomes of the train set are displayed in Table S2, with D-CyPre exhibiting exceptionally high Precision values for several enzymes among the nine CYP450 enzymes in both the train set and test set. For instance, the Precision values for 2A6 and 2E1 in the training and test sets surpassed 0.8 and 0.9, respectively.

Regrettably, D-CyPre in Precision Mode does not exhibit strong performance across all enzymes. Despite the fact that D-CyPre performs well for 2B6 and 2C8 in the validation set (Table S2), their results in the test set indicate severe overfitting (Table S4). We observed that CyProduct also encounters this issue, with the models for 2B6 and 2C8 performing well in the validation set but poorly in the test set. Consequently, to further investigate the underlying causes, we employed t-SNE to visualize the SOMs based on features generated by D-MPNN (van der Maaten & Hinton, 2008). We visualize the SOMs in train set, validation set and test set of these models. The green box in Figure 4.A represents potential false negatives in the test set that reduce Recall for the 2B6 model. The part enclosed by the box in Figure 4.B is the possible FN in test set, which reduces the Recall of the model of 2B6. Similarly, Figure 4.C and Figure 4.D illustrate possible sources of error for 2C8. From these results, it can be inferred that there may be two reasons for poor generalization ability of these models on the test set. First, it could be due to an insufficient size of their train sets which leads to some bonds or atoms with similar structures to SOMs in the test set being misclassified as positive while some actual SOMs that are unfamiliar are misclassified as negative.

The second is that the 2B6 and 2C8 having almost the largest Non-SOMs/SOMs (Table S5) in their respective test sets which makes Precision more sensitive to errors, and perhaps the test results of the model will perform more closely to the validation set on larger test sets. Furthermore, we observed that neither test nor validation sets were distributed within regions lacking training data which implies that there were no atoms or bonds present in either set that had not been previously encountered by our models and thus D-CyPre's chemical space based on its training data is sufficiently large.

3.2.2 Recall Mode

As per the results presented in Table S6, in comparison to CyProduct, D-CyPre exhibits a 17% increase in Jaccard Score (WAvg), a 22% increase in Precision (WAvg), a 5% increase in Recall (WAvg), and a 13% increase in F1 (WAvg). Additionally, the Jaccard Scores for 2B6 and 2C8 also improved under Recall Mode. Overall, our models successfully maintained non-low Jaccard Scores while achieving high Recall.

3.3 D-CyPre with the Molecular Features

Initially, we compared the effects of two molecular features on 1A2 and 2B6 (Table S7) and found that molecular features calculated by D-MPNN exhibited some advantages over those calculated using fixed rules. As such, we employed the same methodology to construct a version

of D-CyPre that incorporates molecular features calculated by D-MPNN. However, this version of D-CyPre did not exhibit better performance across all isoforms when compared with the original version of D-CyPre (Table S8 and S9). Subsequently, we synthesized optimal models from both versions of D-CyPre (with or without molecular features) to obtain new Precision Mode (Table 3) and Recall Mode (Table 4). Among them, 1A2, 2A6, 2B6, 2C8, 2C9 and 2C19 enzymes were all ultimately adopted by models incorporating molecular features under both modes which suggests that molecular structure may be an important factor affecting metabolic reactions for these enzymes. In Precision Mode, compared with the Random Predictor and the CyProduct, the D-CyPre increased Jaccard Score by 590% and 18%, Precision by 845% and 43%, and F1 by 393% and 13%. In Recall Mode, compared with the two models, D-CyPre increased Jaccard Score by 603% and 20%, Precision by 727% and 25%, Recall by 40% and 5%, and F1 by 399% and 15%. The parameters for loss function and XGBOOST for all models can be found in Table S10. Finally, the findings suggest that the molecular features is necessary to consider.

CONCLUSIONS

This study proposes a novel SOMs identification tool called D-CyPre. This model is the pioneer of applying D-MPNN to *in silico* metabolism prediction and has achieved satisfactory results with high Precision, Recall and Jaccard Score. D-CyPre comprises a feature generator and SOMs discriminators and is divided into Precision Mode and Recall Mode. Under both modes, the model ensures good Jaccard Scores while maintaining Precision and Recall values greater than 0.7 respectively. As such, D-CyPre's two modes make it better suited to meet the needs of various types of work. For example, when conducting a high-throughput study, we may prefer more accurate results whereas when making predictions for several drugs and comparing their corresponding metabolites' mass spectra we may prefer to consider all possibilities. Additionally, the results indicate that the molecular features is necessary to consider in *in silico* metabolism prediction.

To use the software (supplementary files 3), users simply input a table containing the SMILES of all target compounds. We believe that the model is sophisticated enough to distinguish most similar SOMs from non-SOMs and can be further trained on larger datasets to achieve higher Jaccard Scores and generalization capabilities. Also, it is possible to attempt the development of a generalized approach for predicting the molecular SOMs of various metabolic enzymes based on the ideas presented in this study.

REFERENCES

de Bruyn Kops C, Šícho M, Mazzolari A, Kirchmair J. 2021. GLORYx: Prediction of the Metabolites Resulting from Phase 1 and Phase 2 Biotransformations of Xenobiotics. *Chemical Research in Toxicology* 34:286–299. DOI: 10.1021/acs.chemrestox.0c00224.

de Bruyn Kops C, Stork C, Šícho M, Kochev N, Svozil D, Jeliazkova N, Kirchmair J. 2019. GLORY: Generator of the Structures of Likely Cytochrome P450 Metabolites Based on Predicted Sites of Metabolism. *Frontiers in Chemistry* 7:402. DOI: 10.3389/fchem.2019.00402.

Chen T, Guestrin C. 2016. XGBoost: A Scalable Tree Boosting System. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 785–794. DOI: 10.1145/2939672.2939785.

Chen C, Shi H, Jiang Z, Salhi A, Chen R, Cui X, Yu B. 2021. DNN-DTIs: Improved drug-target interactions prediction using XGBoost feature selection and deep neural network. *Computers in Biology and Medicine* 136:104676. DOI: 10.1016/j.compbiomed.2021.104676.

Dang NL, Matlock MK, Hughes TB, Swamidass SJ. 2020. The Metabolic Rainbow: Deep Learning Phase I Metabolism in Five Colors. *J. Chem. Inf. Model.*:19.

Deng D, Chen X, Zhang R, Lei Z, Wang X, Zhou F. 2021. XGraphBoost: Extracting Graph Neural Network-Based Features for a Better Prediction of Molecular Properties. *Journal of Chemical Information and Modeling* 61:2697–2705. DOI: 10.1021/acs.jcim.0c01489.

Gilmer J, Schoenholz SS, Riley PF, Vinyals O, Dahl GE. 2017. Neural Message Passing for Quantum Chemistry.

Han X, Jia M, Chang Y, Li Y, Wu S. 2022. Directed message passing neural network (D-MPNN) with graph edge attention (GEA) for property prediction of biofuel-relevant species. *Energy and AI* 10:100201. DOI: 10.1016/j.egyai.2022.100201.

He S, Li M, Ye X, Wang H, Yu W, He W, Wang Y, Qiao Y. 2016. Site of metabolism prediction for oxidation reactions mediated by oxidoreductases based on chemical bond. *Bioinformatics*:btw617. DOI: 10.1093/bioinformatics/btw617.

Hou T, Wang J, Zhang W, Xu X. 2007. ADME Evaluation in Drug Discovery. 6. Can Oral Bioavailability in Humans Be Effectively Predicted by Simple Molecular Property-Based Rules? *Journal of Chemical Information and Modeling* 47:460–463. DOI: 10.1021/ci6003515.

Ioffe S, Szegedy C. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift.

Jianing L, Severin T. S, Joseph B, Ramy F, Richard A. F. 2011. IDSite: An Accurate Approach to Predict P450-Mediated Drug Metabolism. *Journal of Chemical Theory and Computation* 7:3829–3845. DOI: https://doi.org/10.1021/ct200462q.

Jin W, Stokes JM, Eastman RT, Itkin Z, Zakharov AV, Collins JJ, Jaakkola TS, Barzilay R. 2021. Deep learning identifies synergistic drug combinations for treating COVID-19. *Proceedings of the National Academy of Sciences* 118:e2105070118. DOI: 10.1073/pnas.2105070118.

Jo J, Kwak B, Choi H-S, Yoon S. 2020. The message passing neural networks for chemical property prediction on SMILES. *Methods* 179:65–72. DOI: 10.1016/j.ymeth.2020.05.009.

van der Maaten L, Hinton H. 2008. Viualizing data using t-SNE. *Journal of Machine Learning Research* 9:2579–2605.

Nebert DW, Russell DW. 2002. Clinical importance of the cytochromes P450. *The Lancet* 360:1155–1162. DOI: 10.1016/S0140-6736(02)11203-7.

354 Porokhin V, Liu L-P, Hassoun S. 2023. Using graph neural networks for site-of-metabolism
355 prediction and its applications to ranking promiscuous enzymatic products.
356 *Bioinformatics* 39:btad089. DOI: 10.1093/bioinformatics/btad089.
357 Šícho M, de Bruyn Kops C, Stork C, Svozil D, Kirchmair J. 2017. FAME 2: Simple and
358 Effective Machine Learning Model of Cytochrome P450 Regioselectivity. *Journal of*
359 *Chemical Information and Modeling* 57:1832–1846. DOI: 10.1021/acs.jcim.7b00250.
360 Šícho M, Stork C, Mazzolari A, de Bruyn Kops C, Pedretti A, Testa B, Vistoli G, Svozil D,
361 Kirchmair J. 2019a. FAME 3: Predicting the Sites of Metabolism in Synthetic
362 Compounds and Natural Products for Phase 1 and Phase 2 Metabolic Enzymes. *Journal*
363 *of Chemical Information and Modeling* 59:3400–3412. DOI: 10.1021/acs.jcim.9b00376.
364 Šícho M, Stork C, Mazzolari A, de Bruyn Kops C, Pedretti A, Testa B, Vistoli G, Svozil D,
365 Kirchmair J. 2019b. FAME 3: Predicting the Sites of Metabolism in Synthetic
366 Compounds and Natural Products for Phase 1 and Phase 2 Metabolic Enzymes. *Journal*
367 *of Chemical Information and Modeling* 59:3400–3412. DOI: 10.1021/acs.jcim.9b00376.
368 Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A Simple Way
369 to Prevent Neural Networks from Overfitting. :30.
370 Stokes JM, Yang K, Swanson K, Jin W, Cubillos-Ruiz A, Donghia NM, MacNair CR, French S,
371 Carfrae LA, Bloom-Ackermann Z, Tran VM, Chiappino-Pepe A, Badran AH, Andrews
372 IW, Chory EJ, Church GM, Brown ED, Jaakkola TS, Barzilay R, Collins JJ. 2020. A
373 Deep Learning Approach to Antibiotic Discovery. *Cell* 180:688-702.e13. DOI:
374 10.1016/j.cell.2020.01.021.
375 Tian S, Cao X, Greiner R, Li C, Guo A, Wishart DS. 2021. CyProduct: A Software Tool for
376 Accurately Predicting the Byproducts of Human Cytochrome P450 Metabolism. *J. Chem.*
377 *Inf. Model.*:13.
378 Tian S, Djoumbou-Feunang Y, Greiner R, Wishart DS. 2018. CypReact: A Software Tool for in
379 Silico Reactant Prediction for Human Cytochrome P450 Enzymes. *Journal of Chemical*
380 *Information and Modeling* 58:1282–1291. DOI: 10.1021/acs.jcim.8b00035.
381 Wang J, Hou T. 2015. Advances in computationally modeling human oral bioavailability.
382 *Advanced Drug Delivery Reviews* 86:11–16. DOI: 10.1016/j.addr.2015.01.001.
383 Xu B, Wang N, Chen T, Li M. 2015. Empirical Evaluation of Rectified Activations in
384 Convolutional Network.
385 Xu X, Zhang W, Huang C, Li Y, Yu H, Wang Y, Duan J, Ling Y. 2012. A Novel Chemometric
386 Method for the Prediction of Human Oral Bioavailability. *International Journal of*
387 *Molecular Sciences* 13:6964–6982. DOI: 10.3390/ijms13066964.
388 Yang K, Swanson K, Jin W, Coley C, Eiden P, Gao H, Guzman-Perez A, Hopper T, Kelley B,
389 Mathea M, Palmer A, Settels V, Jaakkola T, Jensen K, Barzilay R. 2019. Analyzing
390 Learned Molecular Representations for Property Prediction. *Journal of Chemical*
391 *Information and Modeling* 59:3370–3388. DOI: 10.1021/acs.jcim.9b00237.

Yu J, Shi S, Zhang F, Chen G, Cao M. 2019. PredGly: predicting lysine glycation sites for Homo sapiens based on XGboost feature optimization. *Bioinformatics* 35:2749–2756. DOI: 10.1093/bioinformatics/bty1043.

Zanger UM, Schwab M. 2013. Cytochrome P450 enzymes in drug metabolism: Regulation of gene expression, enzyme activities, and impact of genetic variation. *Pharmacology & Therapeutics* 138:103–141. DOI: 10.1016/j.pharmthera.2012.12.007.

Zaretzki J, Matlock M, Swamidass SJ. 2013. XenoSite: Accurately Predicting CYP-Mediated Sites of Metabolism with Neural Networks. *Journal of Chemical Information and Modeling* 53:3373–3383. DOI: 10.1021/ci400518g.

Zhang C, Hu D, Yang T. 2022. Anomaly detection and diagnosis for wind turbines using long short-term memory-based stacked denoising autoencoders and XGBoost. *Reliability Engineering & System Safety* 222:108445. DOI: 10.1016/j.ress.2022.108445.

Table 1(on next page)

Distribution of SOMs for nine CYP450 Isoforms in Data Sets.

1 **Table 1.** Distribution of SOMs for nine CYP450 Isoforms in Data Sets.

Data set	type	1A2	2A6	2B6	2C8	2C9	2C19	2D6	2E1	3A4
EBoMD	Reactants	279	109	149	147	237	221	282	144	474
	SOMs	1847	615	830	906	1372	1368	1685	863	3139
	Non-SOMs	18760	5951	9914	11322	17481	16387	21596	7458	43597
EBoMD2	Reactants	16	10	11	9	13	13	24	10	41
	SOMs	64	49	31	49	64	51	158	48	236
	Non-SOMs	1182	631	596	946	1134	1180	2581	258	3788

2

Table 2(on next page)

Descriptors of atom and bond.

1 **Table 2.** Descriptors of atom and bond.

Atom Descriptors	Bond Descriptors
Atomic Number	Bond Type (Single/Double/Triple/Aromatic)
Degree	Conjugation
Formal Charge	Ring Membership
Chirality	Stereochemistry
Number Of Bonded Hydrogens	(-)
Hybridization	(-)
Aromaticity	(-)
Ring Membership	(-)
Atomic Mass	(-)

2

Table 3(on next page)

Training results (Precision Mode) for nine CYP450 enzymes in EBoMD and EBoMD2.

a: The results of D-CyPre on train set; b: The results of D-CyPre on validation set; c: The results of D-CyPre on EBoMD; d: The results of D-CyPre on EBoMD2; e: The microaverage (weighted average, weighted by the number of SOMs) over the nine.

Table 3. Training results (Precision Mode) for nine CYP450 enzymes in EBoMD and EBoMD2.

	1A2	2A6	2B6	2C8	2C9	2C19	2D6	2E1	3A4	WAvg[e]
	Jaccard Score TP/(TP+FP+FN)									
D-CyPre[a]	0.845	0.919	0.625	0.832	0.680	0.650	0.545	0.728	0.791	0.733
D-CyPre-val[b]	0.475	0.695	0.489	0.500	0.512	0.573	0.550	0.703	0.469	0.527
D-CyPre-all[c]	0.826	0.880	0.644	0.760	0.685	0.660	0.560	0.744	0.765	0.722
D-CyPre-test[d]	0.593	0.549	0.333	0.281	0.639	0.492	0.548	0.469	0.462	0.497
	Precision TP/(TP+FP)									
D-CyPre	0.989	0.994	0.896	0.990	0.831	0.721	0.747	0.860	0.968	0.891
D-CyPre-val	0.832	0.953	0.830	0.729	0.758	0.662	0.843	0.867	0.769	0.792
D-CyPre-all	0.978	0.998	0.702	0.983	0.745	0.751	0.759	0.876	0.962	0.871
D-CyPre-test	0.699	0.933	0.500	0.667	0.852	0.750	0.735	0.958	0.676	0.737
	Recall TP/(TP+FN)									
D-CyPre	0.854	0.924	0.674	0.839	0.790	0.869	0.668	0.825	0.812	0.802
D-CyPre-val	0.526	0.719	0.543	0.614	0.613	0.811	0.613	0.787	0.546	0.618
D-CyPre-all	0.841	0.881	0.887	0.770	0.896	0.844	0.682	0.832	0.788	0.812
D-CyPre-test	0.797	0.571	0.500	0.327	0.719	0.588	0.684	0.479	0.593	0.610
	F1 2 × Precision × Recall/(Precision+Recall)									
D-CyPre	0.917	0.958	0.769	0.908	0.810	0.788	0.705	0.842	0.883	0.841
D-CyPre-val	0.645	0.820	0.657	0.667	0.678	0.729	0.710	0.825	0.639	0.688
D-CyPre-all	0.904	0.936	0.784	0.864	0.814	0.795	0.718	0.853	0.866	0.835
D-CyPre-test	0.745	0.708	0.500	0.439	0.780	0.659	0.709	0.639	0.632	0.660

a: The results of D-CyPre on train set; b: The results of D-CyPre on validation set; c: The results of D-CyPre on EBoMD; d: The results of D-CyPre on EBoMD2; e: The microaverage (weighted average, weighted by the number of SOMs) over the nine.

Table 4(on next page)

Training results (Recall Mode) for nine CYP450 enzymes in EBoMD and EBoMD2.

a: The results of D-CyPre on train set; b: The results of D-CyPre on validation set; c: The results of D-CyPre on EBoMD; d: The results of D-CyPre on EBoMD2; e: The microaverage (weighted average, weighted by the number of SOMs) over the nine.

Table 4. Training results (Recall Mode) for nine CYP450 enzymes in EBoMD and EBoMD2.

	1A2	2A6	2B6	2C8	2C9	2C19	2D6	2E1	3A4	WAvg[e]
	Jaccard Score TP/(TP+FP+FN)									
D-CyPre[a]	0.872	0.915	0.644	0.788	0.688	0.835	0.575	0.729	0.646	0.723
D-CyPre-val[b]	0.501	0.708	0.577	0.504	0.554	0.619	0.561	0.709	0.517	0.561
D-CyPre-all[c]	0.842	0.907	0.644	0.774	0.685	0.829	0.588	0.742	0.656	0.722
D-CyPre-test[d]	0.571	0.636	0.358	0.365	0.580	0.463	0.554	0.500	0.468	0.506
	Precision TP/(TP+FP)									
D-CyPre	0.970	0.965	0.689	0.848	0.752	0.923	0.636	0.820	0.728	0.799
D-CyPre-val	0.798	0.934	0.652	0.702	0.678	0.748	0.664	0.830	0.657	0.719
D-CyPre-all	0.957	0.956	0.702	0.840	0.745	0.920	0.643	0.837	0.751	0.803
D-CyPre-test	0.658	0.854	0.463	0.519	0.734	0.660	0.615	1.000	0.570	0.645
	Recall TP/(TP+FN)									
D-CyPre	0.897	0.946	0.907	0.918	0.890	0.897	0.857	0.868	0.852	0.882
D-CypPe-val	0.574	0.746	0.833	0.641	0.752	0.781	0.783	0.830	0.708	0.725
D-CyPre-all	0.875	0.946	0.887	0.907	0.896	0.894	0.872	0.868	0.838	0.876
D-CyPre-test	0.813	0.714	0.613	0.551	0.734	0.608	0.848	0.500	0.725	0.720
	F1 2 × Precision × Recall/(Precision+Recall)									
D-CyPre	0.932	0.955	0.783	0.882	0.815	0.910	0.730	0.843	0.785	0.835
D-CyPre-val	0.668	0.829	0.731	0.670	0.713	0.764	0.719	0.830	0.682	0.717
D-CyPre-all	0.914	0.951	0.784	0.872	0.814	0.907	0.740	0.852	0.792	0.835
D-CyPre-test	0.727	0.778	0.528	0.535	0.734	0.633	0.713	0.667	0.638	0.669

a: The results of D-CyPre on train set; b: The results of D-CyPre on validation set; c: The results of D-CyPre on EBoMD; d: The results of D-CyPre on EBoMD2; e: The microaverage (weighted average, weighted by the number of SOMs) over the nine.

Figure 1

Overview of D-CyPre Metabolism Prediction suite (shown for a specific instance of CYP2A6).

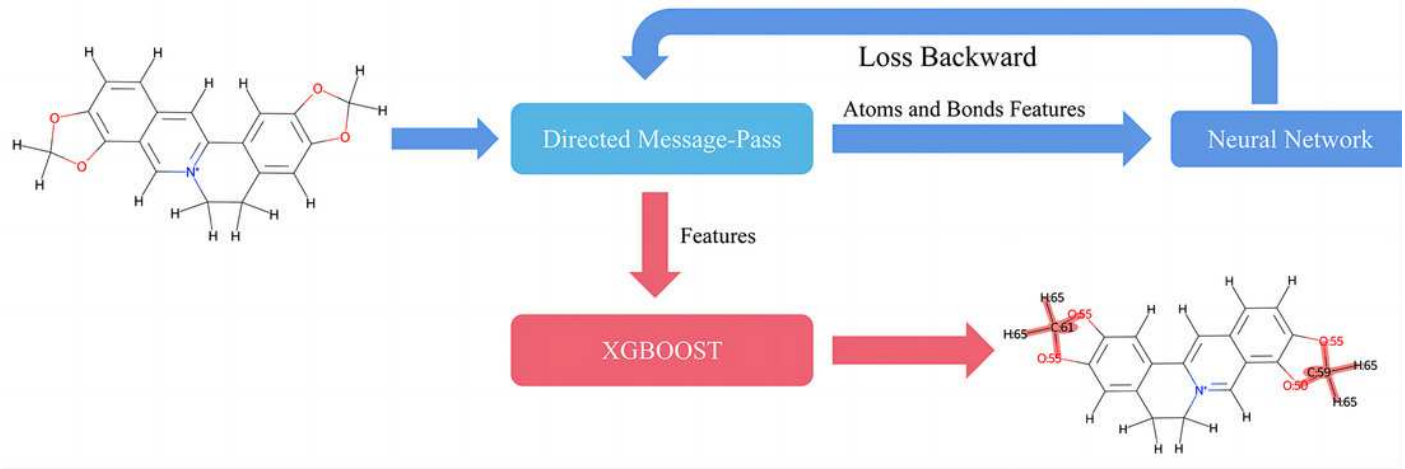


Figure 2

Illustration of our proposed D-CyPre.

D-CyPre employs two independent message passing processes to capture features of two kinds of directed bonds from a molecule. It then fuses the features of the two kinds of directed bonds to derive features of atoms, chemical bonds, and molecules. The features of atoms and bonds are separately combined with those of the molecule and input into a feed forward layer to generate prediction probabilities, which in turn update the network. Moreover, the concatenated features of atoms and bonds are fed into the XGBOOST model to obtain the actual prediction probabilities.

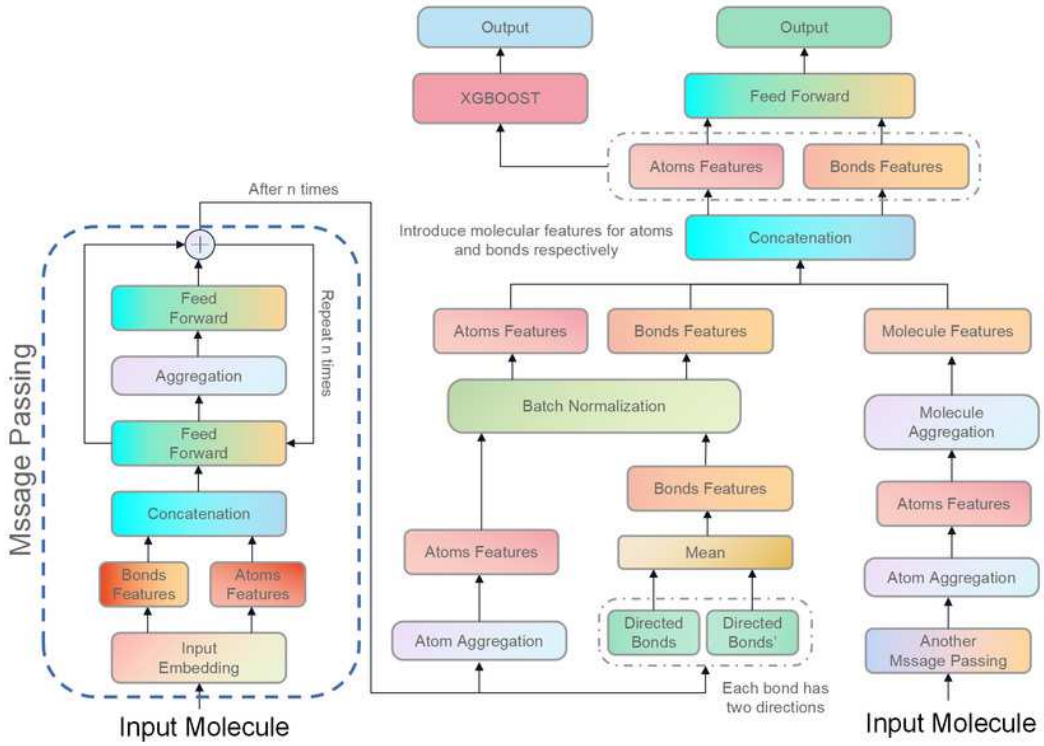


Figure 3

Overview of training model (shown for a specific instance of CYP1A2).

When adjusting parameters, we only use train set (80% of EBoMD) of 1A2 and all data sets (100% of EBoMD) of the other isoforms to train the model. All train set (100% of EBoMD) of 1A2 and the other isoforms (100% of EBoMD) will be used when training the final model.

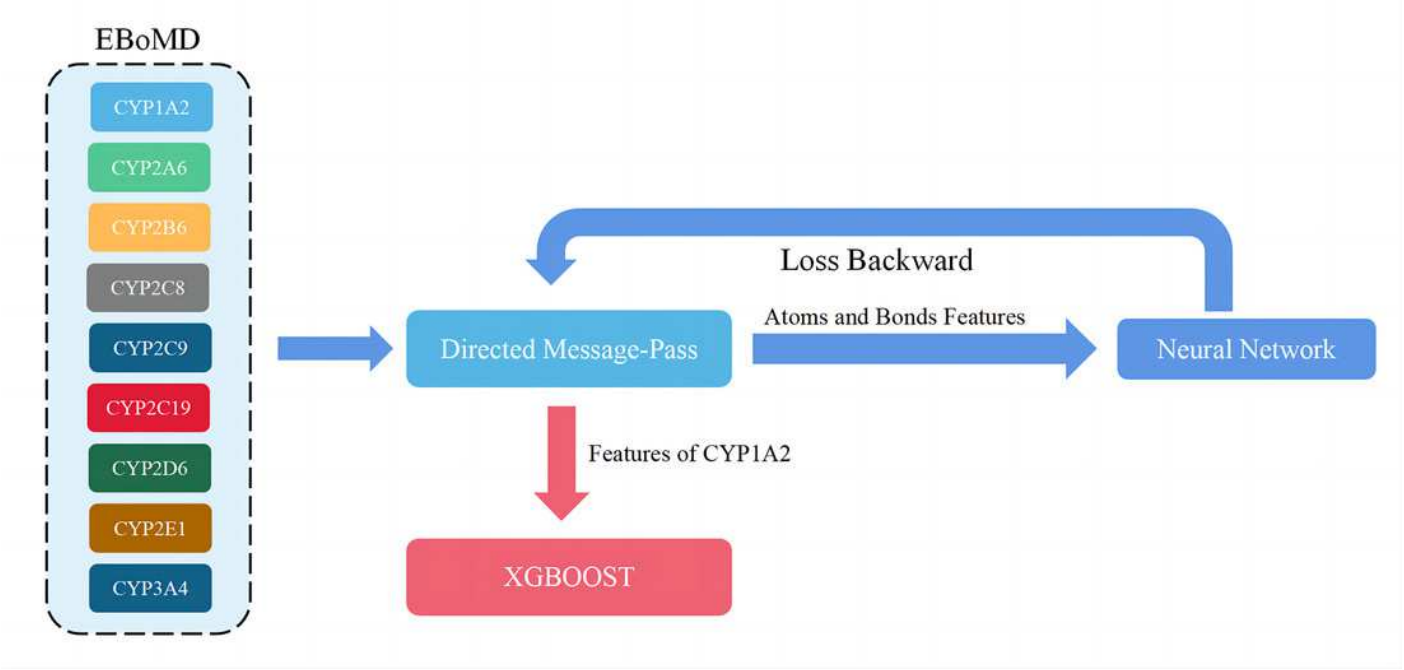
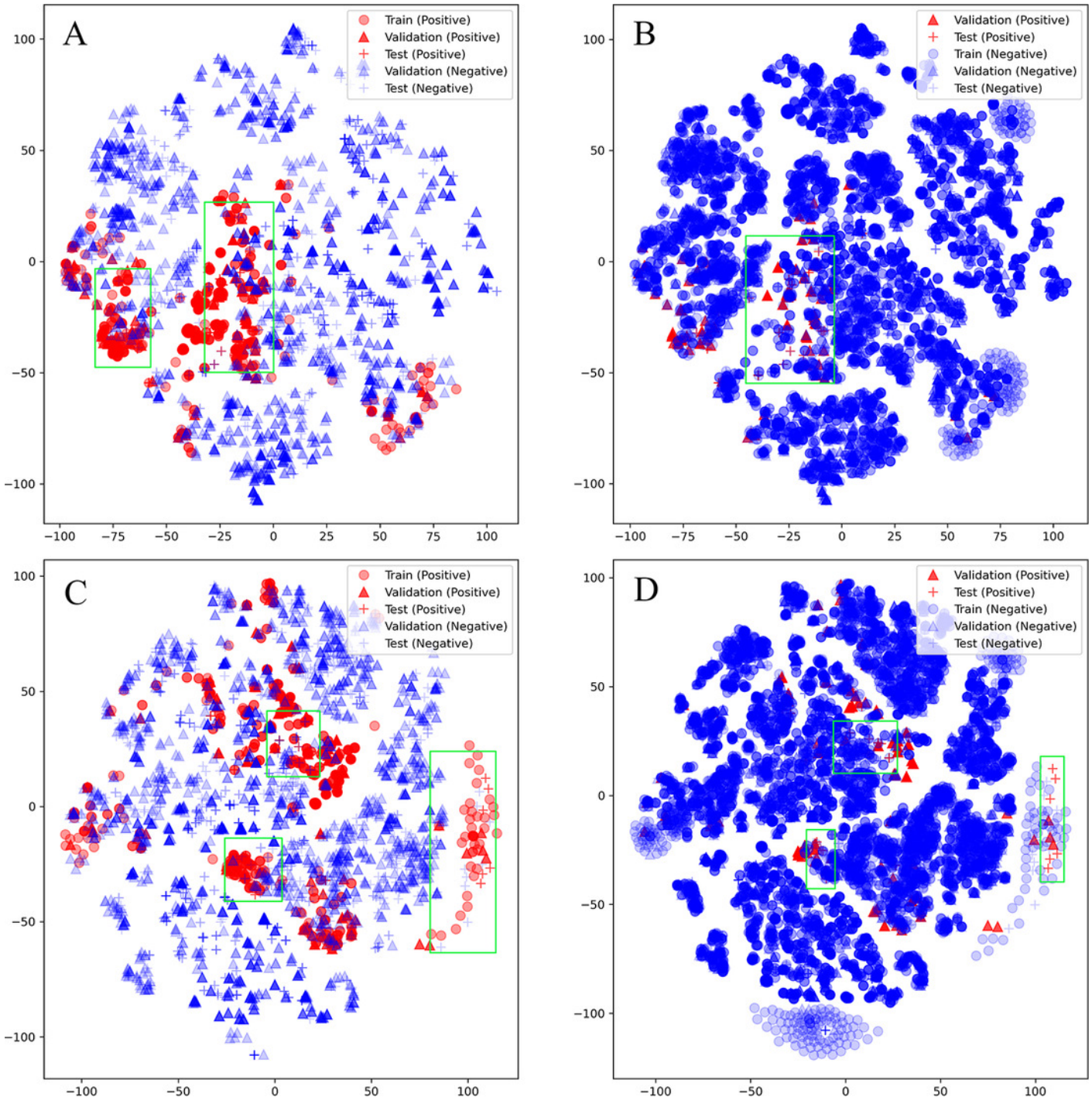


Figure 4

Visualize (by t-SNE) the SOMs of 2B6 and 2C8.

Visualize (by t-SNE) the SOMs of 2B6 (ignore Train; Negative) (A), 2B6 (ignore Train; Positive) (B), 2C8 (ignore Train; Negative) (C) and 2C8 (ignore Train; Positive) (D). The green box part is some data that the model may misjudge.


A
Train (Positive)
Validation (Positive)
Test (Positive)
Validation (Negative)
Test (Negative)
B
Validation (Positive)
Test (Positive)
Train (Negative)
Validation (Negative)
Test (Negative)
C
Train (Positive)
Validation (Positive)
Test (Positive)
Validation (Negative)
Test (Negative)
D
Validation (Positive)
Test (Positive)
Train (Negative)
Validation (Negative)
Test (Negative)