CypEGAT: A Deep Learning Framework Integrating Protein Language Model and Graph Attention Networks for Enhanced CYP450s Substrate Prediction

Yao Wei[1], Uliano Guerrini[1], and Ivano Eberini[1]

Dipartimento di Scienze Farmacologiche e Biomolecolari "Rodolfo Paoletti", Università degli Studi di Milano
yao.wei & uliano.guerrini & ivano.eberini@unimi.it

Abstract. Human Cytochrome P450 enzymes (CYP450s) are responsible for metabolizing 70%-80% of clinically used drugs. The development of computational tools to accurately predict CYP450 enzyme-substrate interactions is crucial for drug discovery and chemical toxicology studies. In this work, we introduce CypEGAT, a deep learning framework designed to enhance prediction performance by integrating protein embeddings of CYP450s (extracted using the pre-trained ESM-2 Transformer model) with molecular embeddings generated by our fine-tuned Graph Attention Network (GAT). The CypEGAT model was trained end-to-end on two large-scale experimental enzyme-substrate datasets and our CYP450s dataset, which comprises 51,753 CYP450 enzyme-substrate pairs and 27,857 enzyme-nonsubstrate pairs. Focusing on five major human CYP450 isoforms (CYP1A2, CYP2C9, CYP2C19, CYP2D6, and CYP3A4), CypEGAT achieves an overall predictive accuracy of 0.882 and an AUROC of 0.928. The model demonstrates robust generalizability to novel chemical compounds across different CYP450 isoforms, underscoring its potential as a powerful tool for drug metabolism studies.

Keywords: Enzyme-substrate prediction · Deep learning · Drug discovery

1 Introduction

Cytochrome P450s (CYP450s), a highly diverse superfamily of heme-thiolate proteins, are indispensable components of the oxidative metabolic machinery found across various life forms. In humans, 57 distinct CYP450 isoforms have been identified, collectively responsible for metabolizing 70-80% of clinically used drugs [4]. Computational approaches to predict interactions between chemical compounds and CYP450s offer significant advantages, such as reducing economic and labor costs, alleviating environmental pollution, and facilitating the preselection of hit compounds for drug discovery and toxicology studies, thus accelerating research progress.

Machine learning algorithms are central to the development of such predictive models. These include support vector machines (SVMs), random forest (RF) [17], [9], deep neural networks (DNNs) [8], and others. However, the performance of these predictive models is often constrained by the quality and quantity of training data, as well as the comprehensiveness of the chemical descriptors used. Addressing these limitations is essential for improving the accuracy and reliability of computational predictions [19].
To overcome these challenges, we have developed a novel deep learning-based model that effectively predicts substrates for five major human CYP450 isoforms (CYP1A2, CYP2C9, CYP2C19, CYP2D6, and CYP3A4). Our main contributions to the existing prediction models are as follows:

- **enhanced deep learning framework:** we propose an improved deep learning approach for CYP450 substrate prediction by fine-tuning a molecular GAT that dynamically updates and aggregates features based on molecular graph structures. Additionally, a feature fusion strategy is employed to integrate protein embeddings;
- **robust pre-training strategy:** by pre-training on three diverse enzyme-substrate datasets, our model demonstrates robust performance across multiple CYP450 isoforms, providing a unified approach to CYP450 substrate prediction.

2 Related Work

Traditional machine learning models to predict CYP450 substrates rely primarily on molecular information, often necessitating separate models for each CYP450 isoform. Recently, there has been a growing interest in enzyme-substrate prediction models that integrate both protein and molecular representations through deep learning techniques. This integration has been facilitated by advances in protein language models (PLMs), such as the ESM Transformer and its variants, for protein representations, and by graph neural networks (GNNs) for molecular representations.
A notable example is ESP [14], which was trained on an extensive experimental dataset of enzyme-substrate pairs from the UniProt-Gene Ontology Annotation (GOA) database. The ESP model utilized a slightly modified ESM-1b Transformer to encode protein embeddings and used a GNN for molecular representations. To train the model, datasets of enzyme-substrate and enzyme-nonsubstrate pairs were generated, and a gradient boosting model was used, achieving high accuracy in predicting novel enzyme-substrate pairs.
Building upon this work, Du et al. introduced FusionESP [6], an enhanced enzyme-substrate predictive model. FusionESP employs a contrastive multimodal fusion strategy that combines protein embeddings encoded by the ESM-2 Transformer with molecular representations generated by MolFormer [20]. These advancements highlight the potential of deep learning models to enhance enzyme-substrate prediction tasks by leveraging both protein and molecular representations, thereby significantly improving predictive performance.

3 Methodology

3.1 Model Architecture

In this study, we present an enhanced deep learning framework, designed to predict substrates for five key human CYP450 isoforms. The model architecture is illustrated in Figure 1.

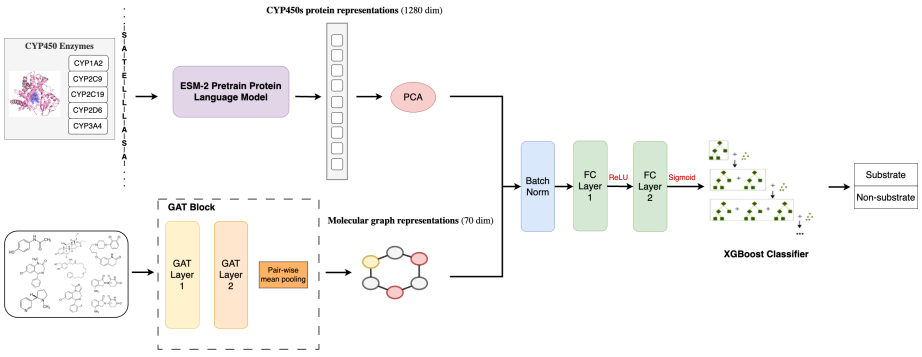

Fig. 1. Overview of CypEGAT Model Architecture. The ESM-2 Transformer encodes CYP450 protein representations, while a modified GAT generates molecular graph representations. These representations are then fused, and a XGBoost classifier is employed to predict CYP450s substrate and non-substrate.

The model incorporates the pre-trained ESM-2 Transformer [15] to encode CYP450s protein representations and employs a fine-tuned GAT to generate molecular representations. Protein embeddings derived from the ESM-2 model were dimensionally reduced using Principal Component Analysis (PCA) and subsequently fused with molecular graph representations. These fused features are passed through two fully connected layers to create a unified enzyme-substrate representation for classification.
In the classification module, we utilize an XGBoost classifier to determine whether a given compound is a substrate or non-substrate for a specific CYP450 isoform. The XGBoost model processes the enzyme-substrate pair representations and outputs a binary classification for each isoform. To enhance performance, we performed five-fold cross-validation to identify the optimal hyperparameters for the XGBoost models.

3.2 Model Graph Attention Network

Graphs provide an intuitive way to represent molecular structures, where nodes correspond to atoms and edges represent chemical bonds. Inspired by the excellent work of Veličković et al. [18], we have developed an enhanced molecular

GAT model for generating molecular graph representations. The architecture of our GAT is illustrated in Figure 2. It consists of two GAT layers: the first layer learns combined features of atoms and bonds, which are then updated and aggregated. These combined features are concatenated with atom-specific features and passed to the second GAT layer. Finally, a feature fusion step constructs enzyme-substrate pairs for classification tasks. Our molecular GAT employs multi-head attention, adjacency masking, and dropout for robust learning and predicts final outputs using fully connected layers.

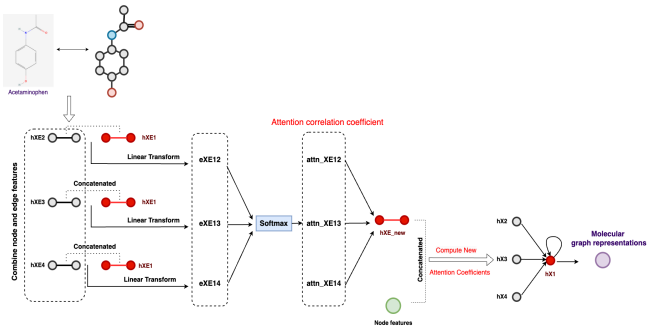

Fig. 2. Diagram of our Molecular GAT.

Let the input node (atom) features be $X \in R^{N \times F_{\text{atom}}}$, the edge (combined atom-bond) features be $XE \in R^{N \times N \times (F_{\text{atom}+F_{bond}})}$. Here, N represents the number of nodes. F_{atom} is the number of features that describe each atom, and F_{bond} is the number of features describing each bond. The adjacency matrix is denoted as $A \in R^{N \times N}$.
In the first GAT layer, a linear transformation is applied to the edge features:

$$H_0^{XE} = XE \cdot W_{\text{edge}} \tag{1}$$

Where, $H_0^{XE} \in R^{N \times N \times (D \cdot H)}$, D is the dimensionality of the feature space for the attention mechanism, and H is the number of attention heads.
Subsequently, the attention coefficient scores for the neighboring edges are calculated:

$$e^{XE} = ReLU(e_i^{XE} + e_j^{XE}) \tag{2}$$

Among them, e_i^{XE} represents the attention scores for atoms, and e_j^{XE} sums the attention coefficients from neighboring edges.
The attention coefficient scores are then masked and normalized:

$$e^{XE} = \text{where}(A > 0, e^{XE}, -\infty),\ \alpha^{XE} = \text{softmax}(e^{XE}) \tag{3}$$

Edge features aggregation is performed as follows:

$$H_{\text{attn}}^{XE} = \alpha^{XE} \cdot H_0^{XE} \tag{4}$$

In the second GAT layer, node and edge features are aggregated:

$$X^1 = \sum_{j \in \mathcal{N}(i)} (H^{XE}_{\text{attn}} + X^1_{ij},\ X^1_{ij} = H^{XE}_{\text{attn}} \cdot W_{m1}) \tag{5}$$

Node attention scores are computed as:

$$e^X_i = ReLU(\text{attn}_i(h^X_i) + \text{attn}_j(h^X_j)) \tag{6}$$

Attention scores are masked and normalized, with dropout applied to prevent overfitting:

$$\alpha^X_{ij} = \text{softmax}(e^X_{ij}) \tag{7}$$

Node features are aggregated using attention scores:

$$H^X_{attn} = \sum_j \alpha^X_{ij} H^X_{ij} \tag{8}$$

The final read out for classification tasks is computed as:

$$h = \text{Concat}(\text{MeanPool}(H_{\text{final}}\ , \text{ESM-2}) \tag{9}$$

$$h1 = ReLU(\text{Linear}(h) \tag{10}$$

$$y = \text{Sigmoid}(\text{Linear}(h_2)) \tag{11}$$

3.3 Dataset

Our curated CYP450s dataset comprises 51,753 enzyme-substrate pairs and 27,857 enzyme-nonsubstrate pairs across five human CYP450 isoforms: CYP1A2, CYP2C9, CYP2C19, CYP2D6, and CYP3A4. These data were collected from studies conducted by Chang et al. [3], Fang et al. [7] and Ai et al. [1].

As illustrated in the scaffold diversity curve (Figure 3, left), the curve exhibits a

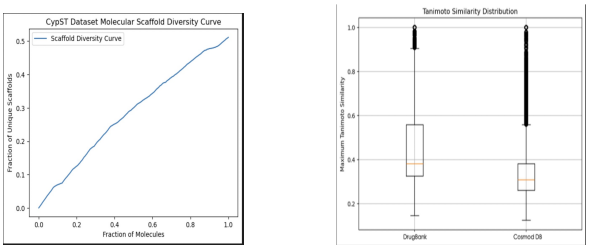

Fig. 3. Left: CypEGAT Dataset Molecular Scaffold Diversity Curve. **Right:** CypEGAT Dataset Molecular Tanimoto Similarity Distribution.

gradual slope, indicating that the molecules in our dataset are chemically diverse

and evenly distributed across various scaffolds. Additionally, Tanimoto similarity analysis reveals that our dataset covers approximately 40% of small-molecule drugs in the DrugBank database [13] and 30% of organic compounds used in cosmetics from the COSMOS DB [21] (Figure 3, right). Each molecule is labeled as either "1" (substrate) or "0" (nonsubstrate), based on its bioactivity data.

3.4 Featurization

To enhance the compatibility and effectiveness of our model, we employed distinct featurization strategies tailored for the CYP450s and molecules.
For the CYP450s, we utilized the pre-trained ESM-2 Transformer model (ESM-2_t33_650M_UR50D) to extract the protein embeddings. In particular, the protein sequences were processed through the ESM-2 model, and the 1280-dimensional embeddings were obtained from the final hidden layer. To reduce the dimensionality of these embeddings and align them with the lower-dimensional molecular representations, we applied Principal Component Analysis (PCA). The dimensionality was determined by the cumulative explained variance [10]. Mathematically, the cumulative explained variance ratio for the first k principal components is defined as:

$$\text{Cumulative Explained Variance Ratio}_k = \frac{\sum_{i=1} k\lambda_i}{\sum_{i=1} n\lambda_i} \tag{12}$$

where λ_i are the eigenvalue of the covariance matrix $\sum$ derived from the data. k represents the number of principal components considered, and n is the total number of principal components (or features). Our analysis (Figure 4) demonstrated that retaining the top 50 principal components captures 99% of the cumulative explained variance, effectively preserving the essential protein features. Thus, we retained these 50 components for subsequent integration with the molecular embeddings.
For the molecular data, we used our fine-tuned GAT model to generate the rep-

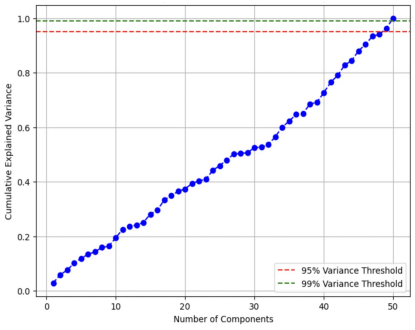

Fig. 4. Cumulative Explained Variances of the CYP450s Protein Representations Principal Components

resentations. In addition, we incorporated customized molecular features that describe key physical and chemical properties, including atom type, bond type, conjugation, aromaticity, ring structures, hydrogen bonding, hybridization, chirality, stereochemistry, and charge. These features were systematically encoded and normalized, as required, to ensure consistency and compatibility across the different input datasets.

3.5 Model Training

Our model was trained to minimize the mean squared error (MSE) loss between the predicted and true representations of the enzyme-substrate pairs. We used MSE loss because it effectively captures the continuous and regression-like nature of our prediction targets, allowing the model to optimize smoothly across a range of prediction values. For future work, we may explore other loss functions (such as binary cross entropy loss) to assess its potential benefits for this task.
The training process utilized the Adam optimizer [12] with a learning rate of 0.1 and a dropout rate of 0.2 to mitigate overfitting. The training was conducted in three stages to progressively enhance the model's performance through pre-training and fine-tuning.

- **Stage 1: pre-training on KEGG Drug Dataset.** The model was initially pre-trained on the KEGG drug dataset [11], which comprises 8,392 enzyme-drug pairs. This stage aimed to provide the model with a fundamental understanding of enzyme-drug interactions, leveraging the large and diverse dataset to initialize robust model weights.
- **Stage 2: training on ESP Experiment Dataset.** Following the first pre-training, the model was trained on the ESP experimental dataset, which contains 18,351 enzyme-substrate pairs sourced from the GOA database [5]. This step further refined the model understanding by exposing it to a broader range of enzyme-substrate interactions, reinforcing generalization across biologically relevant patterns.
- **Stage 3: fine-tuning on our CYP450s Dataset.** Finally, the pre-trained model was fine-tuned on our CYP450s dataset. Fine-tuning on the target-specific dataset allowed the model to specialize its predictions for the CYP450 enzyme family, which is the main focus of our study.

For each stage, the datasets were split into training and test sets in an 80:20 ratio. Each set was normalized and divided into mini-batches of size 64. The model was trained for a total of 20 epochs across the three datasets, with training and validation losses computed after each mini-batch to monitor performance and select the best model weights.
The progressive training strategy was designed to maximize the benefits of transfer learning. Pre-training on larger and more general datasets (KEGG and ESP) optimized the model weights to capture broad interaction patterns, enabling better performance during fine-tuning on the smaller target-specific CYP450s dataset.

The results demonstrated the effectiveness of this approach. At the end of the training, the training and validation losses for the KEGG dataset reached 0.9122, while for the ESP dataset, they reached 0.2076. For our CYP450s dataset, the training and validation losses were 0.1912. These outcomes indicate that the model was well-trained, with no evidence of overfitting, and benefited from the multi-stage training pipeline.

4 Results

4.1 Comparison of Different Methods for Model Performance

In this study, we evaluated the impact of different methods on the performance of the CypEGAT model. We utilized several ESM Transformers (ESM-1b [2], ESM-1b-ts, ESM-2) to generate CYP450s protein representations. For molecular representations, we compared Molecular Extended Connectivity Fingerprints (ECFPs), Graph Neural Network (GNN), and our Graph Attention Network (GAT). Additionally, we also assessed the performance of two machine learning classifiers: XGBoost and Multi-Layer Perceptron (MLP). Among these methods, ESM-1b-ts and GNN refer to the ESP model's methodology. The results are shown in Table 1.

Among the different combinations, the best-performing model was ESM-2 +

Table 1. Accuracy, AUROC and 95% Confidence Interval of Accuracy values for different protein and molecular representations, classifiers trained on the CyEGAT model using the CYP450s dataset

Protein	**Molecule**	**Classifier**	**ACC**	**AUROC**	**95% CI for ACC**
ESM-1b	ECFP	MLP	0.818	0.882	(88.59, 89.18)
ESM-1b	ECFP	XGBoost	0.805	0.899	(89.44, 90.78)
ESM-1b	GNN	MLP	0.825	0.843	(85.04, 86.37))
ESM-1b	GNN	XGBoost	0.840	0.895	(89.98, 91.23)
ESM-1b	GAT	MLP	0.831	0.847	(84.52, 85.19)
ESM-1b	GAT	XGBoost	0.842	0.887	(90.04, 91.68)
ESM-1b-ts	ECFP	MLP	0.819	0.887	(85.11, 86.44)
ESM-1b-ts	ECFP	XGBoost	0.810	0.895	(89.23, 90.51)
ESM-1b-ts	GNN	MLP	0.840	0.845	(85.90, 87.28)
ESM-1b-ts	GNN	XGBoost	0.855	0.910	(89.57, 90.91)
ESM-1b-ts	GAT	MLP	0.851	0.849	(86.20, 87.55)
ESM-1b-ts	GAT	XGBoost	0.869	0.921	(89.02, 90.13)
ESM-2	ECFP	MLP	0.816	0.898	(84.37, 85.43)
ESM-2	ECFP	XGBoost	0.822	0.886	(90.62, 91.78)
ESM-2	GNN	MLP	0.839	0.851	(85.46, 86.74)
ESM-2	GNN	XGBoost	0.870	0.910	(90.14, 91.52)
ESM-2	GAT	MLP	0.843	0.870	(86.62, 87.81)
ESM-2	GAT	XGBoost	0.882	0.928	(90.70, 91.83)

GAT + XGBoost, which outperformed other configurations. ESM-2 showed particularly high accuracy when used with GNN and GAT molecular representations. Among the molecular encodings, GNN and GAT yielded competitive results, especially when combined with XGBoost and ESM-2.
Beyond molecular representations, the choice of protein representations and classifiers significantly impacted on the performance. Unlike ESM-1b, which uses absolute sinusoidal positional encoding, ESM-2 integrates Rotary Position Embedding (RoPE). RoPE enables the model to extrapolate beyond its training context by applying relative position encoding. This is achieved by multiplying query and key vectors with sinusoidal embeddings in the self-attention mechanism [16]. The ability to capture relative positional information likely makes ESM-2 protein representations more compatible with GAT molecular representations.

4.2 Prediction Ability on Individual CYP450 Isoform

We evaluated the performance of the CypEGAT model to predict substrates of individual CYP450 isoforms. For comparison, we also assessed the performance of three other CYP450 substrate prediction models: CypReat[17], CYPstrate [9], and MTL [7], which only consider molecular information. The source codes of CypReach and CYPstrate have not been published, so - for a fair comparison - we trained our model on the same datasets used by these models and compared the AUROC results with those reported in their original publications. Additionally, for MTL, we compared the performance of our model to that of the MTL-GAT model, and based the dataset's random partitioning strategy.
The results of our comparative analysis are summarized in Table 2. A summary of the dataset information used for each model is provided in Table 3. We

Table 2. Comparison of Predictive Model AUROC Results on Individual CYP450 isoform

Model	**1A2**	**2C9**	**2C19**	**2D6**	**3A4**
CypReact	0.86	0.83	0.83	0.87	0.92
CypEGAT	0.921	0.907	0.907	0.925	0.934
CYPstrate	0.88	0.87	0.86	0.92	0.92
CypEGAT	0.915	0.920	0.910	0.927	0.929
MTL-GAT	0.929	0.880	0.932	0.912	0.872
CypEGAT	0.937	0.934	0.929	0.937	0.930

also evaluated the performance of the ESP model on our dataset to compare its performance results with those of CypEGAT. The AUROC results are shown in Table 4.
In comparison to the CYP450s substrate prediction models, that rely solely on molecular representations, CypEGAT demonstrated competitive performance in predicting substrates for various CYP450 isoforms. Moreover, when compared

Table 3. CYP450s Substrate Prediction Models Dataset Information (Training/Test)

Model	**1A2**	**2C9**	**2C19**	**2D6**	**3A4**
CypReact	1632/124	1632/128	1632/120	1632/121	1632/132
CYPstrate	1380/346	1378/346	1379/346	1384/347	1408/353
MTL	1364/226	1769/248	1296/182	1982/257	3007/379

Table 4. Comparison of ESP Model AUROC Results on CypEGAT Dataset

Model	**1A2**	**2C9**	**2C19**	**2D6**	**3A4**
ESP	0.892	0.883	0.887	0.901	0.876
CypEGAT	0.919	0.922	0.913	0.932	0.931

to the general enzyme-substrate prediction model ESP, which incorporates both protein and molecular information, CypEGAT outperformed ESP in predicting CYP450 substrate interactions.

5 Conclusion

This paper introduces CypEGAT, an enhanced deep learning framework for predicting CYP450 substrates. Our proposed GAT dynamically updates and aggregates molecular features to generate robust representations. By combining ESM-2 protein embeddings with molecular GAT representations through feature fusion, CypEGAT extracts comprehensive enzyme-molecule information. Pretrained on two general enzyme-substrate datasets and fine-tuned on our CYP450 dataset, CypEGAT demonstrates superior performance in predicting CYP450 substrate interactions.

6 Acknowledgement

This work was supported by the European Union's Horizon Europe research and innovation program under the Marie Skłodowska-Curie grant agreement No. 101073546 (MSCA Doctoral Network Metal-containing Radical Enzymes – MetRaZymes), and grants from MIUR - "Progetto Eccellenza 2023 – 2027".

References

1. Ai, D., Cai, H., Wei, J., Zhao, D., Chen, Y., Wang, L.: Deepcyps: A deep learning platform for enhanced cytochrome p450 activity prediction. Frontiers in Pharmacology **14**, 1099093 (2023)
2. Brandes, N., Goldman, G., Wang, C.H., Ye, C.J., Ntranos, V.: Genome-wide prediction of disease variant effects with a deep protein language model. Nature Genetics **55**(9), 1512–1522 (2023)
3. Chang, J., Fan, X., Tian, B.: Deepp450: Predicting human p450 activities of small molecules by integrating pretrained protein language model and molecular representation. Journal of Chemical Information and Modeling **64**(8), 3149–3160 (2024)

4. Danielson, P.á.: The cytochrome p450 superfamily: biochemistry, evolution and drug metabolism in humans. Current drug metabolism **3**(6), 561–597 (2002)
5. Dimmer, E.C., Huntley, R.P., Alam-Faruque, Y., Sawford, T., O'Donovan, C., Martin, M.J., Bely, B., Browne, P., Mun Chan, W., Eberhardt, R., et al.: The uniprotgo annotation database in 2011. Nucleic acids research **40**(D1), D565–D570 (2012)
6. Du, Z., Fu, W., Guo, X., Caragea, D., Li, Y.: Fusionesp: Improved enzyme-substrate pair prediction by fusing protein and chemical knowledge. bioRxiv pp. 2024–08 (2024)
7. Fang, J., Tang, Y., Gong, C., Huang, Z., Feng, Y., Liu, G., Tang, Y., Li, W.: Prediction of cytochrome p450 substrates using the explainable multitask deep learning models. Chemical Research in Toxicology (2024)
8. Fu, L., Shi, S., Yi, J., Wang, N., He, Y., Wu, Z., Peng, J., Deng, Y., Wang, W., Wu, C., et al.: Admetlab 3.0: an updated comprehensive online admet prediction platform enhanced with broader coverage, improved performance, api functionality and decision support. Nucleic Acids Research p. gkae236 (2024)
9. Holmer, M., de Bruyn Kops, C., Stork, C., Kirchmair, J.: Cypstrate: a set of machine learning models for the accurate classification of cytochrome p450 enzyme substrates and non-substrates. Molecules **26**(15), 4678 (2021)
10. Jolliffe, I.T., Cadima, J.: Principal component analysis: a review and recent developments. Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences **374**(2065), 20150202 (2016)
11. Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M., Hirakawa, M.: Kegg for representation and analysis of molecular networks involving diseases and drugs. Nucleic acids research **38**(suppl_1), D355–D360 (2010)
12. Kingma, D.P.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
13. Knox, C., Wilson, M., Klinger, C.M., Franklin, M., Oler, E., Wilson, A., Pon, A., Cox, J., Chin, N.E., Strawbridge, S.A., et al.: Drugbank 6.0: the drugbank knowledgebase for 2024. Nucleic acids research **52**(D1), D1265–D1275 (2024)
14. Kroll, A., Ranjan, S., Engqvist, M.K., Lercher, M.J.: A general model to predict small molecule substrates of enzymes based on machine and deep learning. Nature communications **14**(1), 2787 (2023)
15. Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., et al.: Evolutionary-scale prediction of atomic-level protein structure with a language model. Science **379**(6637), 1123–1130 (2023)
16. Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., Liu, Y.: Roformer: Enhanced transformer with rotary position embedding. Neurocomputing **568**, 127063 (2024)
17. Tian, S., Djoumbou-Feunang, Y., Greiner, R., Wishart, D.S.: Cypreact: a software tool for in silico reactant prediction for human cytochrome p450 enzymes. Journal of chemical information and modeling **58**(6), 1282–1291 (2018)
18. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y.: Graph attention networks. arXiv preprint arXiv:1710.10903 (2017)
19. Wei, Y., Palazzolo, L., Mariem, O.B., Bianchi, D., Laurenzi, T., Guerrini, U., Eberini, I.: Investigation of in silico studies for cytochrome p450 isoforms specificity. Computational and Structural Biotechnology Journal (2024)
20. Wu, F., Radev, D., Li, S.Z.: Molformer: Motif-based transformer on 3d heterogeneous molecular graphs. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 5312–5320 (2023)
21. Yang, C., Cronin, M., Arvidson, K., Bienfait, B., Enoch, S., Heldreth, B., Hobocienski, B., Muldoon-Jacobs, K., Lan, Y., Madden, J., et al.: Cosmos next generation–a

public knowledge base leveraging chemical and biological data to support the regulatory assessment of chemicals. Computational Toxicology **19**, 100175 (2021)