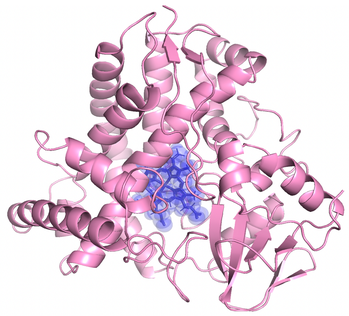
The Use of Machine Learning in Human Cytochrome P450s Substrate Prediction

Yao Wei[1], Uliano Guerrini[1], Ivano Eberini[1]
[1] Department of Pharmacological and Biomolecular Sciences "Rodolfo Paoletti", DiSFeB, University of Milan, Via Balzaretti 9 20133 Milan, Italy
yao.wei@unimi.it

Introduction

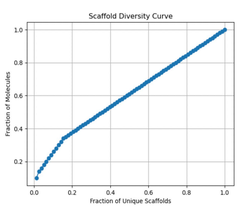
Cytochrome P450s (CYP450s), a highly diverse superfamily of heme-thiolate proteins, are indispensable components of the oxidative metabolic machinery found across various life forms. In humans, 57 distinct CYP450 isoforms have been identified, collectively responsible for metabolizing 70-80% of clinically used drugs. 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.



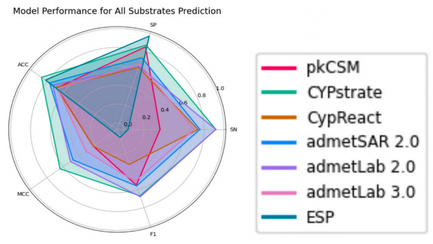
Computational techniques are based on two main approaches: structure-based and ligand-based approaches. The former one relies on the available three-dimensional (3D) protein structures to directly assess the interactions between CYP450s and chemicals. On the other hand, ligand-based approaches can evaluate the structural similarities between ligands and known substrates. The predominant ligand-based method is the QSAR modeling which is used to establish correlations between molecular descriptors of chemicals and their biological activities through mathematical models, thereby explaining the inherent relationships at a molecular level. Nowadays, machine learning algorithms have become the mainstay for building the QSAR models.

Evaluation of Published *In Silico* CYP450s Prediction Tools

100 most prescribed drugs → Testing 7 machine learning prediction models performance → Results



Model	Methods	Year
pkCSM	RF, LR	2015
CypReact	Learning-based model	2018
admetSAR 2.0	RF, k-NN, SVM	2019
CYPstrate	RF, SVM	2021
ADMETlab 2.0	multi-task graph attention	2021
ESP	Esm-1b transformer	2023
ADMETlab 3.0	Multi-task DMPNN	2024



Conclusion:

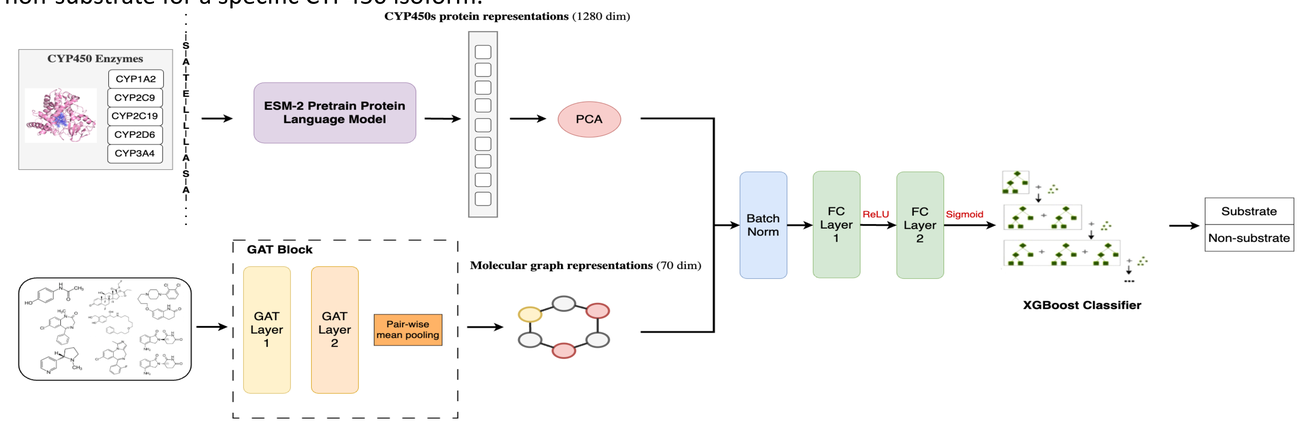
- Both classical machine learning methods and deep learning methods can achieve a certain prediction accuracy.
- Classical machine learning-based models are limited by their datasets scope and computational resources.
- Deep learning-based models are newly developed and models robustness and performance still need to be improved.

Reference:
[1] Wei, Yao, et al. "Investigation of in silico studies for cytochrome P450 isoforms specificity." Computational and Structural Biotechnology Journal (2024).
[2] Wei, Yao, et al. "CypEGAT: A Deep Learning Framework Integrating Protein Language Model and Graph Attention Networks for Enhanced CYP450s Substrate Prediction." AAAI2025, AI4Research workshop.

Funding:
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".

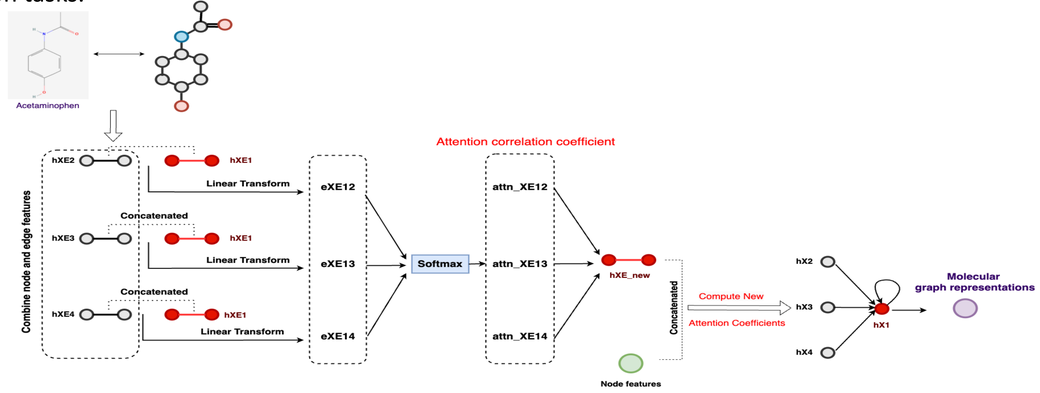
Our CYP450s Substrate Prediction Machine Learning Model: CypEGAT

We present an enhanced deep learning framework, **CypEGAT**, designed to predict substrates for five key human CYP450 isoforms. The model incorporates the pre-trained ESM-2 Transformer to encode CYP450s protein representations and employs a fine-tuned molecular graph attention network to generate molecular representations. 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.



Our Fine-tuned Molecular Graph Attention Network (GAT)

Our molecular GAT employs multi-head attention, adjacency masking, and dropout for robust learning and predicts final outputs using fully connected layers. 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.



Results

Comparison of CypEGAT with Published Prediction Models: **AUROC** Results for Individual Isoforms.

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
ESP	0.892	0.883	0.887	0.901	0.876
CypEGAT	0.919	0.922	0.913	0.932	0.931