chemistry proceedings

MDPI

Proceeding Paper

Reimagining QSAR Modelling with Quantum Chemistry: A CYP1B1 Inhibitor Case Study †

Abanish Biswas and Venkatesan Jayaprakash *

Department of Pharmaceutical Sciences & Technology, Birla Institute of Technology, Mesra, Ranchi 835215, India; email@email.com

* Correspondence: venkatesanj@bitmesra.ac.in

† Presented at the 29th International Electronic Conference on Synthetic Organic Chemistry (ECSOC-29); Available online: https://sciforum.net/event/ecsoc-29.

Abstract

Cytochrome P450 1B1 (CYP1B1) is an important anticancer target due to its overexpression in tumors and role in carcinogen metabolism. In this work we attempted to build a first-generation QSAR model for 63 synthesized inhibitors using Quantum Chemical Descriptors (QCD) and Thermodynamic Descriptors (TCD) derived from xTB calculations. After descriptor reduction was done by multicollinearity analysis and recursive feature elimination (RFE). was built with eight selected descriptors. Validation included k-fold cross-validation, leave-one-out CV, bootstrapping, Y-randomization, and applicability domain analysis. The Among different classifiers cross validation (CV) Support Vector Classifier (rbf kernel) model showed promising internal validation (accuracy ~0.72, ROC-AUC ~0.79), but stringent validations revealed bias toward predicting actives (recall ~1.0, ROC-AUC collapse). Y-randomization confirmed the non-random nature of the structure–activity relationship, while the Williams plot indicated most compounds were within the applicability domain. Although preliminary, this work demonstrates the feasibility of quantum descriptor-based QSAR modeling of CYP1B1 inhibitors and outlines pathways for improving model balance and predictive power

Keywords: CYP1B1 inhibitors; QSAR modeling; quantum chemical descriptors; thermodynamic descriptors; machine learning; cross-validation; Y-randomization; applicability domain

Academic Editor(s): Name

Published: date

Citation: Biswas, A.; Jayaprakash, V. Reimagining QSAR Modelling with Quantum Chemistry: A CYP1B1 Inhibitor Case Study. *Chem. Proc.* **2025**, *volume number*, x. https://doi.org/10.3390/xxxxx

1. Introduction

Cytochrome P450 1B1 (CYP1B1) is a heme-containing enzyme involved in the metabolism of xenobiotics and endogenous substrates. Overexpression of CYP1B1 has been reported in several cancers, including breast, prostate, and ovarian cancer, where it contributes to the activation of procarcinogens into reactive quinones [1]. Selective inhibition of CYP1B1 has therefore emerged as a promising anticancer strategy [2]. Computational modeling plays a vital role in early-stage drug discovery by prioritizing compounds for experimental testing. In particular, Quantitative Structure–Activity Relationship (QSAR) modeling provides a mathematical framework that correlates chemical structure with biological activity [3]. QSAR has traditionally relied on 2D and 3D descriptors such as topological indices, molecular fingerprints, and geometric parameters. While these approaches have been successful, they may not fully capture the electronic, quantum, and

thermodynamic properties that govern enzyme–ligand interactions [4]. Recent studies have highlighted the importance of integrating quantum chemical descriptors (e.g., HOMO, LUMO, energy gaps, molecular charges) and thermodynamic parameters (e.g., enthalpy, entropy, Gibbs free energy) into QSAR models, as they offer deeper physicochemical insights [5]. The semi-empirical extended tight-binding (xTB) method provides a computationally efficient framework for extracting such descriptors, making it suitable for medium-sized QSAR datasets [6]. In this work, we tried to build a first-generation QSAR classification model for CYP1B1 inhibitors using quantum chemical and thermodynamic descriptors derived from xTB calculations. A dataset of 63 compounds synthesized and tested in a single laboratory was employed to ensure biological consistency [7–9]. We evaluated the model using multiple validation approaches, including k-fold cross-validation, leave-one-out CV, Y-randomization, bootstrapping, and applicability domain analysis. Our objective was not only to assess predictive performance but also to critically analyze the limitations and future improvements required for deploying quantum descriptor-based QSAR models.

2. Materials and Methods

The dataset consisted of 63 CYP1B1 inhibitors, for which experimentally determined IC_{50} values were converted into pIC_{50} for uniformity. Molecular descriptors were generated using the xTB framework on Google Colab, which provided a free and reproducible cloud-based computational environment. The descriptors included quantum properties (HOMO, LUMO, energy gap, gradient norm), charge descriptors (maximum, minimum, and range of atomic charges, and sum of absolute charges), and thermodynamic parameters such as zero-point energy (ZPE), enthalpy, entropy, Gibbs free energy, and heat capacity (Cv). To reduce redundancy and multicollinearity, descriptors were subjected to variance inflation factor (VIF) analysis and recursive feature elimination (RFE), yielding a final subset of eight informative descriptors. A Support Vector Classifier (SVC) with a radial basis function kernel was employed for model development. Model performance was rigorously assessed using 10-fold cross-validation, leave-one-out CV (LOOCV), Y-randomization, bootstrapping (100 iterations), and applicability domain analysis using the Williams plot.

.3. Results

The performance of six machine learning classifiers was evaluated using 10-fold cross-validation on the dataset of 63 CYP1B1 inhibitors. Random Forest gave the highest accuracy (0.846) and F1 score (0.889), with an excellent ROC-AUC of 0.9875, followed by Gradient Boosting and XGBoost (accuracy 0.769, F1 0.842). KNN also showed good performance (accuracy 0.769, ROC-AUC 0.913). Logistic Regression achieved lower accuracy (0.692) but maintained perfect recall (1.0). The SVC model displayed the lowest accuracy (0.615), although it produced an ROC-AUC of 1.0, suggesting issues with probability calibration rather than genuine separation.

Cross-Validation Metrics

Table 1 highlights key differences among the classifiers. Ensemble approaches such as Random Forest, Gradient Boosting, and XGBoost consistently outperformed single algorithms, reflecting their ability to reduce variance and capture complex, non-linear relationships. Random Forest emerged as the most balanced model, combining high accuracy and F1 score with excellent ROC-AUC.

Table 1. Performance of classifiers under 10-fold cross-validation.

Model	Accuracy	Precision	Recall	F1	ROC-AUC
Random Forest	0.846	0.800	1.000	0.889	0.9875
Gradient Boosting	0.769	0.727	1.000	0.842	0.9500
XGBoost	0.769	0.727	1.000	0.842	0.8750
KNN (k = 5)	0.769	0.778	0.875	0.824	0.9125
Logistic Regression	0.692	0.667	1.000	0.800	0.7250
SVC (RBF)	0.615	0.615	1.000	0.762	1.0000

Values represent mean performance metrics from 10-fold cross-validation. ROC-AUC = area under the receiver operating characteristic curve. SVC = Support Vector Classifier.

In contrast, Logistic Regression and SVC illustrate the limitations of linear and margin-based classifiers. Logistic Regression achieved perfect recall but sacrificed accuracy and AUC, indicating a strong bias toward predicting the active class. The SVC model gave the lowest accuracy overall, and its apparent ROC-AUC of 1.0 likely reflects calibration artifacts rather than genuine separation.

The KNN classifier performed moderately well, showing that local similarity in descriptor space can partially capture activity trends, but ensemble methods offered superior generalization.

The superiority of ensemble methods is further illustrated in the ROC curves (Figure 1), where Random Forest and Gradient Boosting exhibit the most distinct separation between classes. In addition, the Y-randomization test (Figure 2) confirms that the observed predictive performance is not due to chance: accuracies of randomized models cluster around 0.5, whereas the actual model remains substantially higher, indicating a genuine structure–activity signal.

Figure 3 shows the Williams plot for applicability domain analysis. The majority of compounds fall within the critical leverage threshold and the ±3 standardized residuals boundary, indicating that the model is applicable to most of the dataset. Only a few compounds were identified as potential outliers, suggesting that predictions are generally reliable within the studied chemical space.

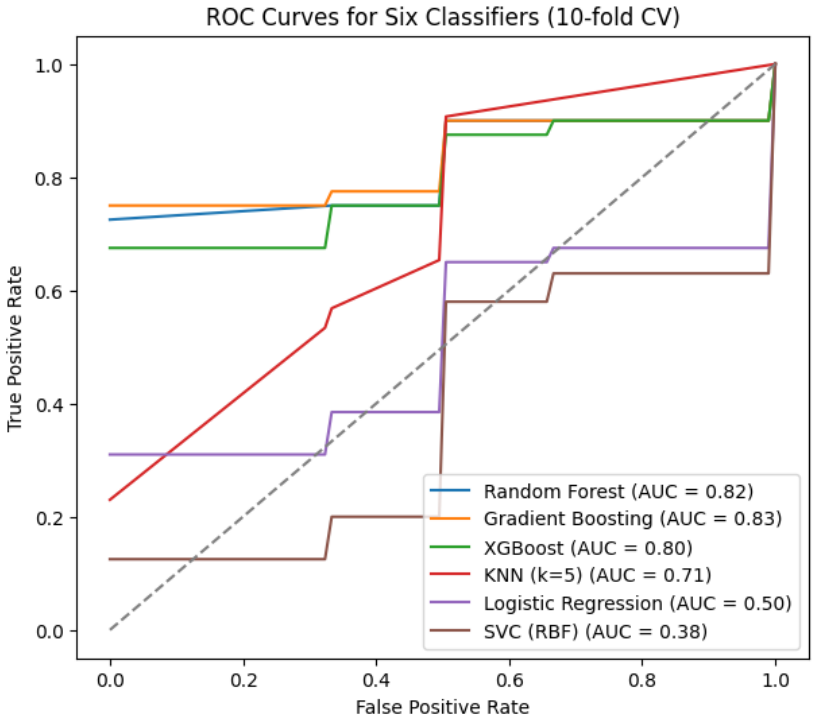

Figure 1. ROC curves for six classifiers under 10-fold cross-validation.

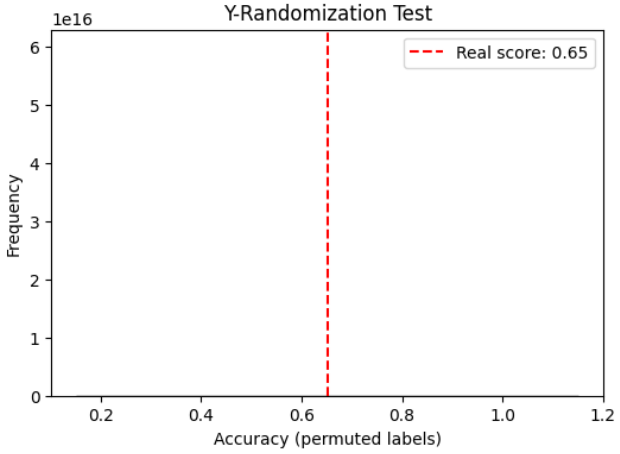

Figure 2. Y-randomization histogram comparing real vs randomized accuracies.

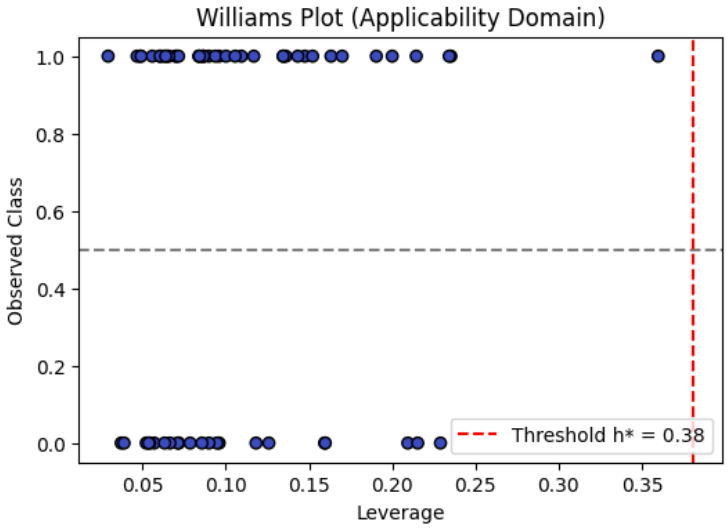

Figure 3. Williams plot to show the applicability domain.

Author Contributions:

Funding: This research was funded by Indian Council of Medical Research.

Institutional Review Board Statement:

Informed Consent Statement:

Data Availability Statement:

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data.

Abbreviations

The following abbreviations are used in this manuscript:

CYP1B1	Cytochrome P450 1B1
QSAR	Quantitative Structure–Activity Relationship
QCD	Quantum Chemical Descriptors
TCD	Thermodynamic Descriptors
xTB	Extended Tight-Binding method
RFE	Recursive Feature Elimination
CV	Cross-Validation
ROC-AUC	Receiver Operating Characteristic – Area Under the Curve
SVC	Support Vector Classifier
RBF	Radial Basis Function (kernel)
KNN	k-Nearest Neighbors
LOOCV	Leave-One-Out Cross-Validation
ZPE	Zero-Point Energy
VIF	Variance Inflation Factor
IC_{50}	Half-maximal inhibitory concentration
pIC_{50}	Negative logarithm of IC50 ($-\log_{10}$IC50)
Cv	Heat Capacity at constant volume

References

1. Murray, G.I.; Melvin, W.T.; Greenlee, W.F.; Burke, M.D. Regulation, Function, and Tissue-Specific Expression of Cytochrome P450 CYP1B1. *Annu. Rev. Pharmacol. Toxicol.* **2001**, *41*, 297–316. https://doi.org/10.1146/annurev.pharmtox.41.1.297.
2. Biswas, A.; Jayaprakash, V. Phytoestrogens and Their Synthetic Analogues as Substrate Mimic Inhibitors of CYP1B1 – An Update (2020–2025). *Bioorganic Med. Chem.* **2025**, *130*, 118385. https://doi.org/10.1016/j.bmc.2025.118385.
3. Cherkasov, A.; Muratov, E.N.; Fourches, D.; Varnek, A.; Baskin, I.I.; Cronin, M.; Dearden, J.; Gramatica, P.; Martin, Y.C.; Todeschini, R.; et al. QSAR Modeling: Where Have You Been? Where Are You Going To? *J. Med. Chem.* **2014**, *57*, 4977–5010. https://doi.org/10.1021/jm4004285.
4. Hansch, Corwin.; Fujita, Toshio. P-σ-π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. *J. Am. Chem. Soc.* **1964**, *86*, 1616–1626. https://doi.org/10.1021/ja01062a035.
5. Gramatica, P. Principles of QSAR Models Validation: Internal and External. *QSAR Comb. Sci.* **2007**, *26*, 694–701. https://doi.org/10.1002/qsar.200610151.
6. Bannwarth, C.; Ehlert, S.; Grimme, S. GFN2-xTB—An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions. *J. Chem. Theory Comput.* **2019**, *15*, 1652–1671. https://doi.org/10.1021/acs.jctc.8b01176.
7. Siddique, M.U.M.; McCann, G.J.; Sonawane, V.; Horley, N.; Williams, I.S.; Joshi, P.; Bharate, S.B.; Jayaprakash, V.; Sinha, B.N.; Chaudhuri, B. Biphenyl Urea Derivatives as Selective CYP1B1 Inhibitors. *Org. Biomol. Chem.* **2016**, *14*, 8931–8936.
8. Siddique, M.U.M.; McCann, G.J.; Sonawane, V.R.; Horley, N.; Gatchie, L.; Joshi, P.; Bharate, S.B.; Jayaprakash, V.; Sinha, B.N.; Chaudhuri, B. Quinazoline Derivatives as Selective CYP1B1 Inhibitors. *Eur. J. Med. Chem.* **2017**, *130*, 320–327.

9. Williams, I.S.; Joshi, P.; Gatchie, L.; Sharma, M.; Satti, N.K.; Vishwakarma, R.A.; Chaudhuri, B.; Bharate, S.B. Synthesis and Biological Evaluation of Pyrrole-Based Chalcones as CYP1 Enzyme Inhibitors, for Possible Prevention of Cancer and Overcoming Cisplatin Resistance. *Bioorganic Med. Chem. Lett.* **2017**, *27*, 3683–3687.