

Optimizing ChEMBL-derived QSAR models for natural flavonoid screening: chemotype-specific predictive reliability in α -glucosidase-inhibition

Radhiah Zakaria¹, Muhammad Iqhrammullah¹, Bryan Gervais de Liyis², Anggi Susilawati³, Nizam Albar⁴, Andriy Anta Kacaribu⁵, Muhammad Habiburrahman⁶

¹Postgraduate Program of Public Health, Universitas Muhammadiyah Aceh, Banda Aceh - Indonesia

²Department of Neurosurgery, National Brain Center Hospital Prof. Dr. Mahar Mardjono, Jakarta - Indonesia

³Faculty of Teacher Training and Education, Universitas Bina Bangsa Getsempena, Banda Aceh - Indonesia

⁴Department of Computer Engineering, Faculty of Engineering, Universitas Serambi Mekkah, Banda Aceh - Indonesia

⁵Doctoral Program of Agricultural Sciences, Postgraduate School, Universitas Syiah Kuala, Banda Aceh - Indonesia

⁶Faculty of Medicine, Imperial College London, London - United Kingdom

ABSTRACT

Introduction: α -Glucosidase inhibitors (AGIs) are essential for controlling postprandial hyperglycemia, with flavonoids representing a major class of natural AGIs. However, conventional quantitative structure–activity relationship (QSAR) models trained on structurally diverse datasets often exhibit limited predictive reliability for natural flavonoids due to chemical space mismatch. This study aimed to develop and optimize Random Forest–based QSAR models derived from ChEMBL AGIs and to evaluate their chemotype-specific applicability to natural flavonoids.

Methods: ChEMBL-derived inhibitors were curated into two datasets: (i) a general scaffold-diverse dataset and (ii) a flavonoid-specific dataset. A custom RDKit-based pipeline employing SMARTS pattern recognition and fingerprint similarity automatically classified flavonoid chemotypes while excluding nonphenolic compounds. Molecular descriptors were calculated using Mordred. Regression and classification models were constructed with 10-fold cross-validation. External validation was performed using 15 literature-reported natural flavonoids categorized into flavones/flavonols (aglycones), flavonol glycosides, isoflavones, and flavan-3-ols.

Results: The general model ($n = 563$) demonstrated strong regression performance ($R^2 = 0.837$) and classification accuracy (0.915). The flavonoid-specific model showed moderate regression ($R^2 = 0.564$) and classification accuracy (0.880). However, external validation revealed superior predictive reliability of the flavonoid-specific model, particularly for catechins (MAE = 0.182) and aglycones (MAE = 0.233). Predictive performance decreased for glycosides and isoflavones.

Conclusion: Chemotype-focused QSAR modeling enhances predictive reliability for natural flavonoids. The optimized flavonoid-specific model is suitable for predicting pIC_{50} of flavones and flavonols but should be applied cautiously to glycosylated or structurally divergent subclasses.

Keywords: Chemotype specificity, Flavonoids, α -Glucosidase inhibitor, Mordred descriptors, QSAR, Random Forest

Introduction

α -Glucosidase is a key enzyme in carbohydrate metabolism, catalyzing the cleavage of terminal α -1,4-linked glucose residues from oligosaccharides and disaccharides (1,2). Its inhibition effectively delays postprandial glucose absorption, forming the pharmacological basis for managing type 2

diabetes mellitus (3). α -Glucosidase inhibitors (AGIs) are mainly used in type 2 diabetes mellitus, where postprandial glucose excursions are prominent, and are not considered standard therapy in type 1 diabetes mellitus. Synthetic AGIs, such as acarbose and miglitol, are clinically established but often limited by gastrointestinal side effects and suboptimal safety profiles (4). Acarbose itself is a natural pseudotetra-saccharide originally derived from *Actinoplanes* species, suggesting the relevance of natural-product-oriented chemical space for AGI discovery and modeling. Consequently, natural flavonoids have attracted significant attention as alternative inhibitors due to their structural diversity, biocompatibility, and multifunctional antioxidant properties (5,6).

Flavonoids constitute one of the most chemically diverse classes of plant polyphenols, encompassing subclasses such

Received: February 19, 2026

Accepted: September 9, 2026

Published online: October 5, 2026

Corresponding author:

Muhammad Habiburrahman

email: h.habiburrahman23@imperial.ac.uk



as flavones, flavonols, isoflavones, and flavan-3-ols. Several members of these groups exhibit notable α -glucosidase inhibition, though their potencies vary considerably depending on substitution patterns, planarity, and glycosylation state (7,8). While experimental screening remains the standard for identifying new inhibitors, *in silico* quantitative structure–activity relationship (QSAR) modeling offers a cost-effective and mechanistically interpretable approach to predict inhibitory potential based on molecular descriptors. However, most existing QSAR models are trained on synthetic or mixed-compound datasets that inadequately represent natural flavonoid scaffolds (5,9), resulting in unreliable extrapolation when applied to complex plant-derived compounds.

Despite the broad pharmacological relevance of flavonoids, naturally occurring flavonoids remain relatively underexplored in QSAR modeling compared with synthetic compound libraries (5,10). This gap may be partly attributed to their substantial structural heterogeneity, including differences in hydroxylation, methoxylation, glycosylation, and substitution patterns across flavones, flavonols, flavanones, isoflavones, and related subclasses (11). Moreover, bioactivity data for natural flavonoids are frequently distributed across heterogeneous experimental systems, targets, and assay conditions, thereby complicating dataset curation and limiting model generalizability. In addition, previous computational screening studies have often been organized around specific plant species (5,9), rather than focusing on flavonoids as a distinct and chemically diverse class of natural compounds. Herein, we hypothesized that a chemotype-restricted QSAR model, trained exclusively on flavonoid-derived compounds, would improve predictive fidelity for natural inhibitors compared to a general model trained on mixed scaffolds.

To address this gap, the present study developed and validated a QSAR model for α -glucosidase inhibition based on ChEMBL-derived datasets, with an additional flavonoid-focused subset designed to enhance compatibility with natural compounds. The study aimed to (i) compare the performance of the general model (all compounds) and the subclass-specific model (flavonoids only), (ii) evaluate their predictive reliability across distinct flavonoid chemotypes, and (iii) delineate the structural domain within which the model remains applicable. Furthermore, by examining patterns of overprediction and underprediction, this work provides insight into how molecular features such as glycosylation and planarity influence model behavior. In a previously developed model for hepatitis C virus inhibitor, atomic properties emerged as the important features (12). Assembled tree models, including Random Forest and LightGBM (13), could capture nonlinear relationships and complex substructure–activity patterns, though the former is more robust to noise and less prone to overfitting (14). Herein, the Mordred descriptor generator was selected because it provides comprehensive molecular characterization while remaining fully compatible with reproducible QSAR workflows. Mordred computes more than 1,800 two-dimensional and three-dimensional descriptors, offering broader descriptor coverage than many alternative open-source tools (15). In addition, its native Python implementation enables seamless integration with RDKit-based preprocessing pipelines,

facilitating transparent data handling, descriptor calculation, and model reproducibility. Unlike several legacy or proprietary descriptor generators, Mordred is fully open-source and actively maintained, making it well-suited for scalable QSAR modeling and external validation (15). The resulting framework offers both a predictive and diagnostic tool for screening potential AGIs from natural flavonoids and for defining chemical boundaries where model predictions can be confidently interpreted.

Methods

Data collection and preprocessing

Experimental α -glucosidase inhibitory data were retrieved from the ChEMBL web interface on October 16, 2025, reflecting the publicly available ChEMBL release at the time of access. Bioactivity records were queried for α -glucosidase (EC 3.2.1.20) and restricted to single-protein targets derived from *Saccharomyces cerevisiae*. Only biochemical enzyme inhibition assays reporting a quantitative IC_{50} endpoint with a direct activity (inhibition) relationship were retained. To ensure endpoint homogeneity suitable for QSAR analysis, cell-based assays, phenotypic screens, binding-only measurements, and non- IC_{50} endpoints (such as % inhibition, K_i , and EC_{50}) were excluded. Records with ambiguous relational qualifiers (including “>”, “<”, or “~”), missing numerical IC_{50} values, or inconsistent unit reporting were removed during data curation. All IC_{50} values (nM) were transformed to pIC_{50} ($-\log_{10}(IC_{50} \times 10^{-9})$) for numerical stability and interpretability. Two datasets were prepared: (1) a dataset consisting of all valid inhibitors regardless of scaffold, and (2) a dataset containing only flavonoid-related scaffolds confirmed through automated and manual curation.

Flavonoid subset classification and curation

Flavonoid-like compounds were extracted from the complete AGI dataset obtained from ChEMBL. A custom RDKit-based Python script was developed to identify compounds containing characteristic flavonoid scaffolds using substructure pattern matching and molecular fingerprint similarity. SMARTS-based structural filters were applied to detect core motifs corresponding to flavones, flavonols, isoflavones, chalcones, and flavanones. Each compound was then assigned a preliminary label based on the presence of these substructures, the number of aromatic rings, carbonyl functionality, and phenolic hydroxyl groups. To enhance classification reliability, Morgan fingerprints (radius = 2, 2048 bits) were computed for each compound and compared against reference flavonoid exemplars, including quercetin-, kaempferol-, luteolin-, genistein-, and catechin-like scaffolds. A combined confidence score, integrating SMARTS-based pattern recognition (60%) and fingerprint similarity (40%), was used to confirm the flavonoid classification. Compounds detected without phenolic –OH groups were excluded. The remaining entries were manually inspected to verify structural integrity and correct classification. The final filtered dataset represented the curated set of flavonoid and non-flavonoid AGIs used for subsequent QSAR model development.

Molecular descriptor calculation

Two-dimensional molecular descriptors were computed using the Mordred descriptor calculator (version 1.2.0) implemented in Python to quantitatively represent the structural and physicochemical properties of each compound (15). All structures were converted from SMILES strings into RDKit molecule objects, and descriptors were calculated while ignoring 3D geometry to ensure consistency across datasets. The full Mordred feature set, encompassing constitutional, topological, electronic, and hydrophobic descriptors, was initially generated, and non-informative or constant-value features were subsequently removed. Any missing or undefined descriptor values were imputed as zero, and all IC_{50} values were transformed to pIC_{50} ($-\log_{10}(IC_{50}$ in molar)) before modeling. The resulting descriptor matrices were used as the independent variables for regression and classification model construction, while the experimental pIC_{50} values served as the dependent variable.

Model development

Two predictive models were developed using the Random Forest algorithm implemented in scikit-learn (v1.5). Techniques such as Random Forest have been recommended as a reliable classifier for health data, including dengue prediction (14). The regression model was designed to predict continuous pIC_{50} values, while the classification model categorized compounds as active ($pIC_{50} \geq 5$) or inactive. Both models were trained on Mordred molecular descriptors generated from curated AGI datasets and optimized using 500 estimators with default hyperparameters. Model robustness was evaluated through 10-fold cross-validation, where regression performance was assessed using the coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE), while classification performance was assessed using accuracy, precision, recall, and F_1 -score. The twenty most influential molecular features were identified from Random Forest feature importance values, and the top-ranking Mordred descriptors were analyzed to interpret the structural and physicochemical factors contributing to α -glucosidase inhibitory activity. The developed QSAR models can be accessed at [Online](#).

External validation

An external validation set comprising 15 naturally occurring flavonoids was assembled from peer-reviewed literature with reported α -glucosidase inhibitory activity (16-18). Mordred descriptors were computed for these compounds and aligned with each model's training feature set. The trained regression and classification models from both Model 1 and Model 2 datasets were applied to predict the pIC_{50} values. To examine structural influence on predictive performance, the external flavonoids were grouped according to their chemotypes, including flavones/flavonols (aglycones), flavonol glycosides (1-2 sugar residues), isoflavones, and flavan-3-ols (catechins). Model performance on the external set was evaluated using standard statistical indicators, including the coefficient of determination (R^2), RMSE, MAE, median absolute error, bias (mean signed error), and

the percentage of compounds predicted within ± 0.3 and ± 0.5 log units of the experimental pIC_{50} values.

Statistical analysis and reproducibility

All data preprocessing, molecular descriptor computation, and QSAR modeling were performed in Python (v3.13) using the RDKit (v2024.3.5) and Mordred (v1.2.0) packages for cheminformatics and molecular descriptor generation. Model training and validation were conducted with scikit-learn (v1.5), and data analysis was supported by pandas (v2.2), NumPy (v1.26), and Matplotlib (v3.9) for statistical computation and visualization. Model performance was assessed using standard statistical metrics, including the coefficient of determination (R^2), RMSE, MAE, median absolute error, bias (mean signed error), accuracy, precision, recall, and F_1 -score. The number of predictions within ± 0.3 and ± 0.5 log units of the experimental pIC_{50} values was also reported to indicate prediction reliability. All analyses were performed on a Windows 11 workstation (Intel® Core™ i7, 32 GB RAM).

Results and Discussion

Internal validation of the trained models

Two Random Forest-based QSAR models were developed to predict α -glucosidase inhibitory activity. Model 1, trained on the complete dataset of 563 compounds representing diverse chemical scaffolds, produced strong predictive performance with $R^2 = 0.837$, RMSE = 0.291, and MAE = 0.173 in 10-fold cross-validation. The corresponding classification model also exhibited robust predictive capability (accuracy = 0.915, Precision = 0.938, Recall = 0.933, $F_1 = 0.936$), reflecting a well-balanced ability to distinguish active and inactive inhibitors. In contrast, Model 2, trained exclusively on 108 curated flavonoid structures, yielded moderate regression accuracy ($R^2 = 0.564$, RMSE = 0.289, MAE = 0.225) but maintained stable classification performance (Accuracy = 0.880, Precision = 0.879, Recall = 0.895, $F_1 = 0.887$). The performance parameters for each model are presented in Table 1.

TABLE 1 - Internal cross-validation performance of the QSAR models

Metric	Model 1 (All compounds)	Model 2 (Flavonoid-only)
Dataset size (n)	563	108
Number of descriptors	1,614	1,444
Regression		
R^2	0.837	0.564
RMSE	0.291	0.289
MAE	0.173	0.225
Classification		
Accuracy	0.915	0.88
Precision	0.938	0.879
Recall	0.933	0.895
F_1 -score	0.936	0.887

Feature importance

Feature importance analysis of Model 1 (trained with a dataset containing diverse types of compounds) revealed that the most influential molecular descriptors were *GGI10*, *SssNH*, *BCUTc-1h*, *PEOE_VSA7*, and *PEOE_VSA13*, as presented in Figure 1a. These top-ranked features collectively captured information related to atomic charge distribution, molecular topology, and hydrogen-bonding potential. The *GGI10* and *BCUTc-1h* descriptors describe global molecular connectivity and eigenvalue-weighted atomic properties, indicating that overall molecular branching and electronic distribution strongly influence α -glucosidase inhibitory potency. *SssNH*, representing the number of secondary amine fragments, suggests that the presence of hydrogen-bond donors plays a critical role in ligand–enzyme interaction. The *PEOE_VSA* and *EState_VSA* descriptors quantify partial charge-weighted surface areas, further emphasizing the contribution of charge separation and polarizability to binding affinity. Descriptors such as *VE1_A*, *SdsN*, and *NdsN* reflect molecular size and topological complexity, suggesting that extended conjugation and aromatic character modulate inhibitory strength. The presence of *SlogP_VSA8* among the top contributors also highlights the importance of lipophilicity balance for molecular recognition at the α -glucosidase active site. Collectively, these descriptors indicate that both electrostatic and topological features govern inhibitory activity across the structurally diverse compound set represented in Model 1.

For Model 2, which was trained exclusively on flavonoid structures, feature importance analysis identified *AATSC0i*, *BCUTi-1h*, *AATS8d*, *AATS8v*, and *AATS7v* as the most influential descriptors (Fig. 1b). The dominance of *AATSC0i* and *BCUTi-1h* highlights the importance of atomic autocorrelation and eigenvalue-weighted descriptors that encode atomic charge and polarizability, suggesting that intramolecular charge distribution and structural symmetry are key determinants of α -glucosidase inhibitory activity within flavonoid scaffolds. The *AATS* (averaged Moreau–Broto autocorrelation) and *ATSC* (centered Broto–Moreau autocorrelation) families capture atom-level property relationships along molecular topology, reflecting how the relative positioning of electronegative and hydrophobic atoms influences inhibitory potency. Descriptors such as *AETA_eta_FL*, *SMR_VSA3*,

and *AMID_N* further indicate that molecular flexibility, refractivity, and the presence of amide or hydrogen-bonding functionalities contribute to binding interactions. The inclusion of *GATS8i* and *C1SP2* among the top-ranked features underscores the significance of extended π -conjugation and planarity—characteristic of flavonoid aglycones—in determining activity. Collectively, these descriptors reveal that α -glucosidase inhibition among flavonoids is primarily governed by electronic distribution, topological autocorrelation, and conjugated surface characteristics, consistent with the structural properties of polyphenolic inhibitors.

Correlation heatmaps of the top 20 Mordred descriptors showing inter-descriptor relationships within Model 1 and Model 2 are presented in Figure 2. The analysis revealed several moderate correlations among charge-related and surface area descriptors, particularly within the *PEOE_VSA* and *EState_VSA* families, reflecting their shared dependence on partial atomic charge distributions. In contrast, descriptors such as *GGI10* and *SssNH*, which represent topological and fragment-based features, showed low correlation with most others, indicating that they provide orthogonal and independent contributions to model prediction. The absence of strong multicollinearity among the most influential features suggests that the Random Forest model effectively captured diverse structural and electronic aspects governing α -glucosidase inhibitory activity. The presence of both topological (such as *GGI10*, *BCUTc-1h*) and electronic descriptors (such as *PEOE_VSA7*, *EState_VSA9*) among the dominant, yet weakly correlated features supports a balanced descriptor composition that enhances the model's interpretability and generalization performance.

In contrast, the correlation heatmap for Model 2, trained exclusively on flavonoid structures, showed denser clustering among autocorrelation and charge-weighted descriptors (Fig. 2a). Descriptors from the *AATS* and *ATSC* families exhibited stronger positive correlations, consistent with the structural homogeneity of flavonoid scaffolds that share similar conjugated aromatic systems and hydroxylation patterns. The co-variation among these descriptors indicates that flavonoid activity is primarily governed by interdependent topological and electronic factors describing charge distribution and molecular planarity. Compared with Model 1, which drew on a wider variety of orthogonal features across chemically diverse scaffolds, Model 2 reflects a more interrelated

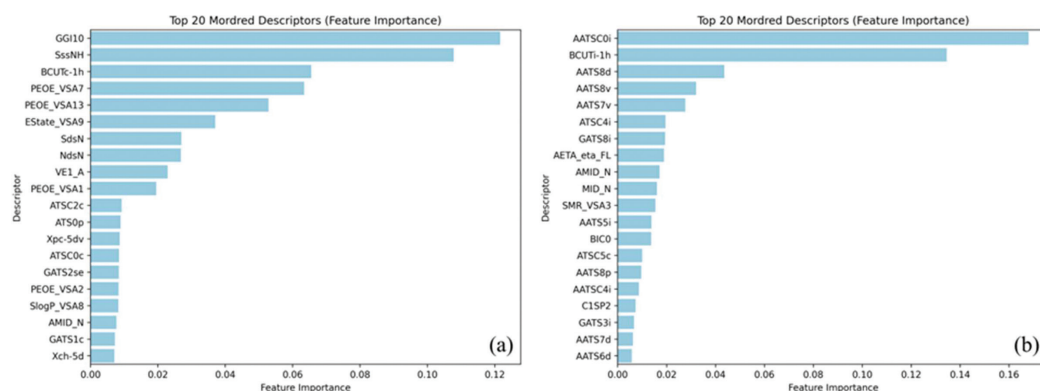


FIGURE 1 - Twenty top important features in Model 1 (a) and Model 2 (b).

descriptor space shaped by the consistent π -conjugated backbone of flavonoids. Together, these observations suggest that while Model 1 captures diverse chemical behavior across broad inhibitor classes, Model 2 emphasizes subtle variations in electronic topology specific to polyphenolic systems.

Chemical space overlap

PCA plots visualizing the chemical space overlap between training and external compounds for both QSAR models are presented in Figure 3. The PCA of Model 1 revealed a broad and dispersed chemical space, with the external flavonoids positioned at the lower and peripheral regions of the distribution. This indicates that while the model encompassed a wide range of AGIs, the flavonoid compounds occupy a distinct subregion, partially outside the model's central applicability domain. Meanwhile, Model 2 demonstrated a more compact and overlapping distribution between training and external compounds, reflecting greater chemical homogeneity and structural consistency. The close clustering observed in Model 2 suggests that its predictions are more domain-relevant and reliable for flavonoid scaffolds.

External validation

External validation using 15 structurally diverse flavonoids demonstrated that both models yielded predictions in close agreement with the experimental pIC_{50} values, where the comparative data are presented in Table 2. The corresponding chemotype-wise validation statistics are presented in Table 3. Model 2 (flavonoid-only) was found to outperform Model 1 (overall compounds) in most categories. For Model 1, the lowest errors were observed in flavan-3-ols and aglycone flavonols ($\text{MAE} < 0.3$), whereas higher deviations occurred in glycosylated flavonols ($\text{MAE} = 0.617$) and isoflavones ($\text{MAE} = 0.505$). In comparison, Model 2 achieved smaller mean and median errors across all groups, particularly for catechins and aglycones ($\text{MAE} = 0.182$ and 0.233 , respectively), with most predictions falling within ± 0.5 log units of experimental values. The improved prediction stability of Model 2 reflects its optimized representation of flavonoid-specific structural and electronic features, whereas Model 1 was trained on a chemically heterogeneous dataset.

Discussion

Comparative analysis of the general and flavonoid-specific models revealed distinct predictive behaviors across flavonoid subclasses. The general model, trained on the complete ChEMBL-derived dataset, identified general inhibitory trends but exhibited reduced precision for structurally complex polyphenols. It consistently underpredicted the potency of highly hydroxylated compounds while overestimating smaller aglycones. These discrepancies likely resulted from descriptor noise introduced by non-flavonoid scaffolds that diluted key structural features governing α -glucosidase inhibition. In contrast, the flavonoid-specific model—trained exclusively on flavonoid scaffolds—demonstrated enhanced numerical stability and chemical domain fidelity. It accurately captured the electronic and topological determinants of inhibitory activity, particularly within planar, conjugated aglycones such as flavones and flavonols.

Despite improved accuracy, systematic deviations emerged in certain subclasses. The model overpredicted the activity of flavonol glycosides, reflecting its limited sensitivity to steric and hydrophilic effects introduced by sugar moieties. Conversely, isoflavones and flavan-3-ols were underpredicted, a consequence of altered aromatic topology and reduced π -conjugation that diminish descriptor correspondence with the trained feature space. These trends highlight the impact of subtle structural perturbations—such as B-ring displacement in isoflavones or C2–C3 bond saturation in catechins—on the model's perception of bioactive motifs. These findings illustrate that QSAR model generalizability depends critically on structural alignment between training and application datasets. The strong performance for flavone and flavonol aglycones underscores the value of domain-specific model development in natural product discovery.

Further, the discrepancies observed for glycosylated flavonols and isoflavones suggest that prediction reliability was influenced by subclass-specific structural features. Glycosylated flavonols were generally overpredicted, likely because sugar moieties introduce additional steric bulk, hydrophilicity, conformational flexibility, and altered hydrogen-bonding capacity that are not fully represented by 2D descriptors. In line with a previous study, the position of the

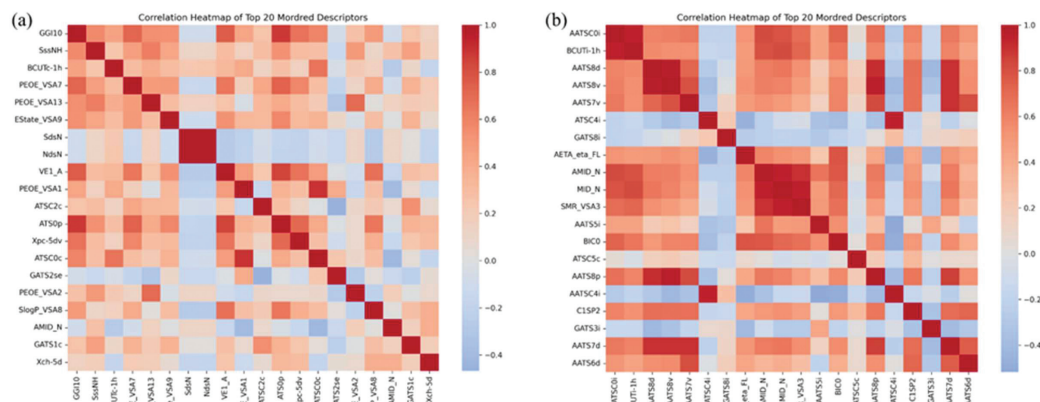


FIGURE 2 - Correlation heatmaps of the top 20 Mordred descriptors used in the Random Forest-based QSAR models 1 (a) and 2 (b). Color intensity from red to blue indicates positive to negative correlation, respectively.

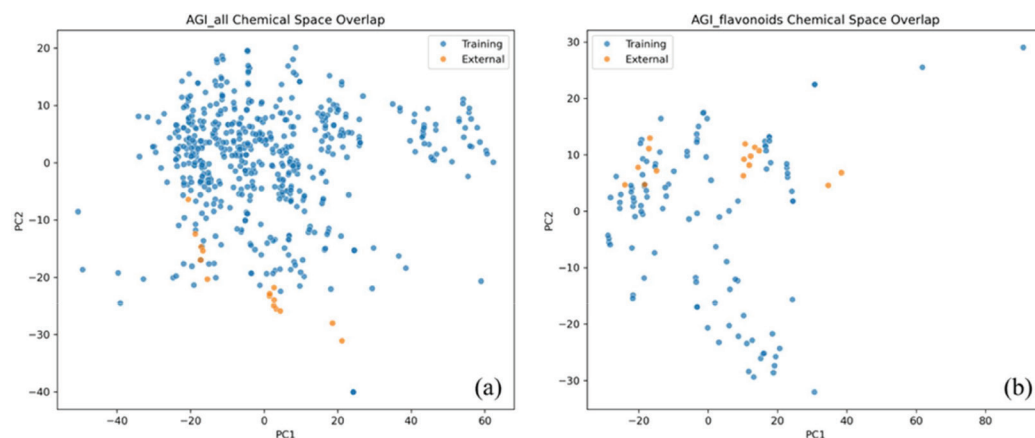
TABLE 2 - Experimental and predicted pIC_{50} for external validation of the QSAR model

Class	Compounds	Experimental pIC_{50}	Predicted pIC_{50}	
			Model 1	Model 2
Flavan-3-ols	(-) Epigallocatechin-gallate	5.28	4.82	4.94
	Epicatechin	4.45	4.43	4.46
	Gallocatechin	4.41	4.57	4.59
Flavones/Flavonols (aglycones)	Cassiaoccidentalin B	4.41	4.94	4.81
	Kaempferol	4.89	4.67	4.70
	Luteolin	4.58	4.63	4.68
Flavonol Glycosides	Astragalin	4.44	4.88	4.77
	Cynaroside	4.68	4.83	4.73
	Isoorientin	4.84	4.82	4.68
	Isoquercetin	3.74	4.94	4.79
	Quercetin-4'-O-glucoside	4.50	4.89	4.69
	Rutin	3.58	5.19	4.90
	Vitexin	4.28	4.78	4.73
Isoflavones	Daidzein	5.14	4.77	4.73
	Genistein	5.28	4.65	4.71

TABLE 3 - Chemotype-wise external validation statistics of predicted pIC_{50} values for α -glucosidase inhibitory activity

Chemotype	n	MAE	RMSE	Median AE	Bias	AE ≤ 0.3 log units	AE ≤ 0.5 log units
Model 1							
Flavan-3-ols	3	0.212	0.283	0.161	-0.105	2	3
Flavones/Flavonols (aglycones)	3	0.269	0.335	0.226	0.119	2	2
Flavonol glycosides	7	0.617	0.816	0.444	0.611	2	4
Isoflavones	2	0.505	0.522	0.505	-0.505	0	1
Model 2							
Flavan-3-ols	3	0.182	0.226	0.183	-0.048	2	3
Flavones/Flavonols (aglycone)	3	0.233	0.264	0.198	0.101	2	3
Flavonol glycosides	7	0.506	0.677	0.33	0.462	3	5
Isoflavones	2	0.494	0.501	0.494	-0.494	0	1

AE, absolute error; RMSE, root mean square error

**FIGURE 3** - PCA plots illustrating the chemical space overlap between training and external compounds for the Random Forest-based QSAR models (a) and (b).

sugar moiety on flavonoid structure affected α -glucosidase inhibition (19,20). These features may reduce membrane or active-site accessibility and shift the compounds away from the aglycone-like chemical space learned by the model proposed in the present study. In contrast, isoflavones tended to be underpredicted, possibly because the altered position of the B-ring changes molecular topology, planarity, and π -conjugation compared with flavones and flavonols (21,22). Thus, the model proposed in the present study performs most reliably for planar aglycone flavonoids but should be applied cautiously to glycosylated or structurally divergent subclasses.

In practical terms, this enables targeted virtual screening of flavonoid scaffolds, where close topological and electronic similarity to the training set ensures quantitative reliability and accurate rank-order correlation. As witnessed previously, hydroxylation patterns (especially C-3, C-4' and B-ring positions), the C2=C3 double bond, and planarity strongly determine AGI potency (23,24). The study concludes that glycosylation generally reduces activity (23). Comparative docking and binding-energy analyses show that the position and presence of glycosylation significantly modulate binding energy (19). This suggests that structural micro-features within the flavonoid scaffold drive activity and that mixing non-flavonoid descriptors can introduce misleading variance. Another published report combining both experimental and *in silico* analyses suggests that the α -glucosidase binding modes are tied to the flavonoid C-ring/A- and B-ring geometry, explaining why flavonoid scaffold features (such as planarity and conjugation) promote consistent interactions with the enzyme active-site pocket (22). As suggested in a previous QSAR analysis, models trained on a coherent scaffold have smaller descriptor variance and more reliable predictions, whereas mixing distant scaffolds widens chemical space (25,26). Thus, a flavonoid-specific model captures these conserved interactions better than a mixed-scaffold model.

The observed biases for glycosides and isoflavones in the present study further expose inherent limitations of conventional 2D descriptor frameworks when chemical variation exceeds the learned structural domain. Glycosylation introduces steric bulk, conformational flexibility, and hydrophilicity not adequately represented by planar descriptors, resulting in systematic overestimation of activity. Predictions for such compounds should therefore be interpreted qualitatively, serving as indicators of inhibitory potential rather than quantitative estimates. This is in line with previous reviews, observing that aglycone forms of many flavonoids show stronger α -glucosidase inhibitory activity than their glycosylated counterparts (7,27). Moreover, changes in hydrogen-bonding pattern and decreased hydrophobic/ π interactions explain the potency loss of AGI candidates (28). Outcomes in the present study, therefore, are consistent with experimental data reporting the effects of glycosylation and reduced planarity on α -glucosidase inhibition (19,21).

Subclass-specific QSAR models augmented with 3D structural descriptors or molecular interaction fingerprints may better capture steric and hydrogen-bonding effects in complex natural products, as previously reported (29,30). The flavonoid-only framework exemplifies how chemical domain

restriction enhances interpretability and predictive precision. By clearly delineating its valid chemotypes, the model provides a foundation for systematic domain expansion and for developing hybrid QSAR-machine learning pipelines that integrate physicochemical, structural, and bioactivity data (31). The model's diagnostic capacity adds further utility, where agreement between predicted and expected activity suggests domain inclusion, while systematic deviation flags compounds requiring additional refinement (32). This interpretative function transforms the model from a predictive algorithm into a mechanistic tool for understanding structure-domain relationships across natural product chemotypes (33). Its adaptability also supports extension to related glycosidase targets, reinforcing its relevance for rational inhibitor design and natural product-based therapeutic discovery.

The major limitation of this study was that all molecular descriptors were calculated from two-dimensional structures without explicit consideration of conformational dynamics or solvent effects, which could influence the α -glucosidase inhibitory activity in biological systems. Therefore, the resulting models may have limited ability to capture structural determinants that depend on three-dimensional molecular geometry, including spatial orientation, intramolecular hydrogen bonding, conformational accessibility, and molecular flexibility. This limitation may be particularly relevant for complex flavonoids, such as glycosylated derivatives, where biological activity may be influenced by steric configuration and target-specific binding interactions. Additionally, the experimental IC_{50} values collected from literature sources may involve variations in assay conditions, potentially introducing systematic bias into model training and validation. Future studies integrating larger, more diverse datasets of natural compounds and incorporating three-dimensional or quantum chemical descriptors could further enhance model generalizability and mechanistic interpretability.

Conclusion

A QSAR model for natural AGI has been successfully developed in this study using ChEMBL data and optimized for flavonoid scaffolds. The refined flavonoid-only model showed high predictive accuracy for planar aglycone structures, particularly flavones and flavonols, where structural overlap between training and natural compound domains was strong. It showed moderate tolerance for catechin-type flavan-3-ols but did not reliably predict activities of glycosylated flavonols and isoflavones. These results demonstrate that domain-focused model design improves interpretability and predictive reliability when extending QSAR frameworks from synthetic to natural compounds. The flavonoid-only model provides a practical tool for *in silico* screening and prioritization of bioactive flavonoids, defining chemotypes with dependable predictions and supporting rational candidate selection and experimental validation.

Acknowledgments

The authors appreciate the inter-institutional collaboration between authors from Universitas Muhammadiyah

Aceh, National Brain Center Hospital Prof. Dr. dr. Mahar Mardjono, Universitas Bina Bangsa Getsempena, Universitas Serambi Mekkah, Imperial College London, and Universitas Syiah Kuala.

Disclosures

Conflict of interest: All the authors declare that there are no conflicts of interest.

Financial support: The research received no external funding.

Data availability: Data underlying this study and the trained models are available on [Online](#).

Authors' contributions: Conceptualization, R.Z., M.I., M.H., and A.A.K.; Methodology, R.Z., M.I., and N.A.; Formal analysis, B.G.dL.; Investigation, R.Z. and B.G.dL.; Data curation, B.G.dL. and N.A.; Validation, N.A. and M.H.; Visualization, A.S.; Resources, N.A.; Writing—original draft preparation, R.Z., B.G.dL., and A.S.; Writing—review & editing, R.Z., M.I., N.A., M.H., and A.A.K.; Supervision, M.I., M.H., and A.A.K. All authors have read and agreed to the published version of the manuscript.

References

- Mushtaq A, Azam U, Mehreen S, et al. Synthetic α -glucosidase inhibitors as promising anti-diabetic agents: recent developments and future challenges. *Eur J Med Chem*. 2023;249:115119. [CrossRef PubMed](#)
- Agrawal N, Sharma M, Singh S, et al. Recent advances of α -glucosidase inhibitors: a comprehensive review. *Curr Top Med Chem*. 2022;22(25):2069–2086. [CrossRef PubMed](#)
- Patel P, Shah D, Bambharoliya T, et al. A review on the development of novel heterocycles as α -glucosidase inhibitors for the treatment of type-2 diabetes mellitus. *Med Chem*. 2024;20(5):503–536. [CrossRef PubMed](#)
- Chen Y, Xiao Y, Lian G, et al. Pneumatosis intestinalis associated with α -glucosidase inhibitors: a pharmacovigilance study of the FDA adverse event reporting system from 2004 to 2022. *Expert Opin Drug Saf*. 2023;1–10. [CrossRef PubMed](#)
- Riyaphan J, Pham DC, Leong MK, et al. In silico approaches to identify polyphenol compounds as α -glucosidase and α -amylase inhibitors against type-II diabetes. *Biomolecules*. 2021;11(12):1877. [CrossRef PubMed](#)
- Benjamin MAZ, Mohd Mokhtar RA, Iqbal M, et al. Medicinal plants of Southeast Asia with anti- α -glucosidase activity as potential source for type-2 diabetes mellitus treatment. *J Ethnopharmacol*. 2024;330:118239. [CrossRef PubMed](#)
- Lam TP, Tran NN, Pham LD, et al. Flavonoids as dual-target inhibitors against α -glucosidase and α -amylase: a systematic review of in vitro studies. *Nat Prod Bioprospect*. 2024;14(1):4. [CrossRef PubMed](#)
- Barber E, Houghton MJ, Williamson G. Flavonoids as human intestinal α -glucosidase inhibitors. *Foods*. 2021;10(8):1939. [CrossRef PubMed](#)
- Diéguez-Santana K, Puris A, Rivera-Borroto OM, et al. A fuzzy system classification approach for QSAR modeling of α -amylase and α -Glucosidase Inhibitors. *Curr Comput Aided Drug Des*. 2022;18(7):469–479. [CrossRef PubMed](#)
- Yang T, Yang Z, Pan F, et al. Construction of an MLR-QSAR model based on dietary flavonoids and screening of natural α -glucosidase inhibitors. *Foods*. 2022;11(24):4046. [CrossRef PubMed](#)
- Amini F, Abbas KI, Ghasemi JB. Molecular modeling approach in design of new scaffold of α -glucosidase inhibitor as antidiabetic drug. *Biochem Biophys Rep*. 2025;42:101995. [CrossRef PubMed](#)
- Noviandy TR, Idroes GM, Maulana A, et al. Optimizing hepatitis C virus inhibitor identification with LightGBM and tree-structured parzen estimator sampling. *Engineering, Technology & Applied Science Research*. 2024;14(6):18810–18817. [CrossRef](#)
- Idroes GM, Noviandy TR, Idroes GM, et al. Prognostication of differentiated thyroid cancer recurrence: an explainable machine learning approach. *Narra X*. 2024;2(3):e183–e183. [CrossRef](#)
- Priyanto D, Hairani H, Marzuki K, et al. Optimization of Random Forest for health data classification using PCA and K-Means SMOTE-ENN. *Engineering, Technology & Applied Science Research*. 2025;15(5):27646–27652. [CrossRef](#)
- Moriwaki H, Tian YS, Kawashita N, et al. Mordred: a molecular descriptor calculator. *J Cheminform*. 2018;10(1):4. [CrossRef PubMed](#)
- Yang D, Wang L, Zhai J, et al. Characterization of antioxidant, α -glucosidase and tyrosinase inhibitors from the rhizomes of *Potentilla anserina* L. and their structure-activity relationship. *Food Chem*. 2021;336:127714. [CrossRef PubMed](#)
- Wang B, Wang C, Duan Y, et al. The effects of *Monascus purpureus* fermentation on metabolic profile, α -glucosidase inhibitory action, and in vitro digestion of mulberry leaves flavonoids. *Lebensm Wiss Technol*. 2023;188:115449. [CrossRef](#)
- Kashtoh H, Baek KH. Recent updates on phytoconstituent α -glucosidase inhibitors: an approach towards the treatment of type two diabetes. *Plants*. 2022;11(20):2722. [CrossRef PubMed](#)
- Dej-adisai S, Sakulkeo O, Wattanapiromsakul C, et al. Flavonoid constituents and α -glucosidase inhibition of *Solanum stramonifolium* Jacq. inflorescence with in vitro and in silico studies. *Molecules*. 2022; 27(23):8189. [CrossRef](#)
- Han X, Wang P, Zhang J, et al. α -Glucosidase inhibition mechanism and anti-hyperglycemic effects of flavonoids from astragalus radix and their mixture effects. *Pharmaceuticals*. 2025; 18(5):744. [CrossRef](#)
- de Souza Farias SA, da Costa KS, Martins JBL. Analysis of conformational, structural, magnetic, and electronic properties related to antioxidant activity: revisiting flavan, anthocyanidin, flavanone, flavanol, isoflavone, flavone, and flavan-3-ol. *ACS Omega*. 2021;6(13):8908–8918 [CrossRef](#)
- Tang H, Huang L, Sun C, et al. Exploring the structure-activity relationship and interaction mechanism of flavonoids and α -glucosidase based on experimental analysis and molecular docking studies. *Food Funct*. 2020; 11 (4): 3332–3350. [CrossRef](#)
- Proença C, Freitas M, Ribeiro D, et al. α -Glucosidase inhibition by flavonoids: an in vitro and in silico structure-activity relationship study. *J Enzyme Inhib Med Chem*. 2017;32(1): 1216–1228. [CrossRef PubMed](#)
- Şöhretoğlu D, Sari S. Flavonoids as α -glucosidase inhibitors: mechanistic approaches merged with enzyme kinetics and molecular modelling. *Phytochem Rev*. 2020;19:1081–1092. [CrossRef](#)
- Li J, Zhao T, Yang Q, et al. A review of quantitative structure-activity relationship: the development and current status of data sets, molecular descriptors and mathematical models. *Chemom Intell Lab Syst*. 2025;256:105278. [CrossRef](#)
- Duke R, Yang CH, Ganapathysubramanian B, et al. Evaluating molecular similarity measures: do similarity measures reflect electronic structure properties? *J Chem Inf Model*. 2025;65(9):4311–4319. [CrossRef PubMed](#)
- Li MM, Chen YT, Ruan JC, et al. Structure-activity relationship of dietary flavonoids on pancreatic lipase. *Curr Res Food Sci*. 2022;6:100424. [CrossRef PubMed](#)

28. Uyanır E, Şoral M, Seyhan G, et al. Alpha-glucosidase inhibitory effects of flavonoids, phenolic acids and iridoids isolated from *Vinca soneri*: in vitro and in silico perspectives. *Chem Biodivers*. 2024;21(10):e202401386. [CrossRef](#)
29. Kızılcan DŞ, Türkmenoğlu B, Güzel Y. Comparison of the performance of different “local reactive descriptors” in 3D-QSAR analysis of enantioselective molecules. *Struct Chem*. 2022;33(2):433-443. [CrossRef](#)
30. Edache EI, Uzairu A, Mamza PA, et al. Structure-based simulated scanning of rheumatoid arthritis inhibitors: 2D-QSAR, 3D-QSAR, docking, molecular dynamics simulation, and lipophilicity indices calculation. *Sci Afr*. 2022;15:e01088. [CrossRef](#)
31. Vasilev B, Atanasova M. A (comprehensive) review of the application of quantitative structure–activity relationship (QSAR) in the prediction of new compounds with anti-breast cancer activity. *Appl Sci (Basel)*. 2025;15(3):1206. [CrossRef](#)
32. De P, Kar S, Ambure P, et al. Prediction reliability of QSAR models: an overview of various validation tools. *Arch Toxicol*. 2022;96(5):1279-1295. [CrossRef PubMed](#)
33. Zhou Y, Li S, Zhao Y, et al. Quantitative structure–activity relationship (QSAR) model for the severity prediction of drug-induced rhabdomyolysis by using random forest. *Chem Res Toxicol*. 2021;34(2):514-521. [CrossRef PubMed](#)