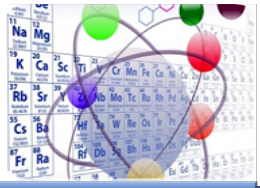
**Indonesian Journal of Chemical Science and Technology (IJCST)**

State University of Medan, https://jurnal.unimed.ac.id/2012/index.php/aromatika

IJCST-UNIMED 2026, Vol. 09, No. 2, Page; 417 – 430

Received : Jul 1st, 2026 Accepted : Jul 25th, 2026 Web Published : Jul 31st, 2026



# Machine Learning–Driven Discovery of Dietary Polyphenol DPP-4 Inhibitors via Molecular Docking

Nur Laily Harfita[1*], Ahmad Faisal Nasution[1], Zuliana Amalia[1], Joceline Schellenberg W[2], Rahayu[3]

[1]Department of Nutrition, Faculty of Health, Universitas Deztron Indonesia, Medan 20233, Indonesia
[2]Department of Informatics, Faculty of Technology and Business, Universitas Deztron Indonesia, Medan 20233, Indonesia
[3]Department of Chemistry, Faculty of Science and Technology, Universitas Samudra, Langsa 24416, Indonesia

*Corresponding author : nurlailyharfita@udi.ac.id

**ABSTRACT**

*T2DM is a chronic metabolic disorder of rising global prevalence, in which DPP-4 serves as a key therapeutic target through its role in incretin hormone degradation. This study aimed to identify polyphenolic compounds derived from dietary sources as potential DPP-4 inhibitors through an in silico pipeline integrating machine learning (ML) and molecular docking. The bioactivity dataset for DPP-4 was retrieved from ChEMBL (CHEMBL284), processed into a binary classification dataset, and represented using 2048-bit Morgan fingerprints combined with five RDKit descriptors. Six ML models were integrated into a soft-voting ensemble, achieving a Matthews correlation coefficient (MCC) of 0.8334 and an AUC-ROC of 0.9739. Screening of 162 compounds from the Phenol-Explorer database yielded 11 potential active inhibitors. Molecular docking identified hesperetin (−8.548 kcal/mol) as the leading candidate, followed by pelargonidin, daidzein, and naringenin, with key binding residues including Ser209, Glu205/206, Tyr631, Arg125, and Tyr662.*

Keywords: Dietary polyphenols; Machine learning; Virtual screening; Molecular docking; Phenol-Explorer

## 1. INTRODUCTION

Type 2 diabetes mellitus (T2DM) represents one of the most pressing global health challenges of the 21st century. According to the International Diabetes Federation, approximately 537 million adults were living with diabetes in 2021, and this number is projected to increase to 783 million by 2045 (International Diabetes Federation [IDF], 2021). T2DM is closely associated with a range of chronic complications, including nephropathy, retinopathy, neuropathy, and cardiovascular diseases, all of which significantly contribute to increased morbidity, mortality, and the global healthcare burden [1].

One of the well-validated therapeutic targets in T2DM management is the enzyme dipeptidyl peptidase-4 (DPP-4). DPP-4 is a transmembrane serine protease that plays a key role in degrading incretin hormones, particularly glucagon-like peptide-1 (GLP-1) and glucose-dependent insulinotropic polypeptide (GIP), thereby



e-ISSN : 2622-4968, p-ISSN : 2622-1349 

limiting their glucose-lowering effects [2]. Inhibition of DPP-4 prolongs incretin activity, enhances glucose-dependent insulin secretion, suppresses glucagon release, and ultimately improves glycemic control [3]. Although DPP-4 inhibitors (gliptins), such as sitagliptin, have demonstrated clinical efficacy, their cost and accessibility remain significant challenges, particularly in developing countries [4,5].

In this context, dietary polyphenols have emerged as promising candidates for safer long-term antidiabetic interventions. Polyphenols are plant-derived secondary metabolites widely distributed in fruits, vegetables, seeds, spices, and beverages, and are known for their diverse biological activities, including antioxidant and antidiabetic effects [6]. Several studies have reported that certain classes of polyphenols, such as flavonoids and phenolic acids, exhibit DPP-4 inhibitory activity in vitro [7,8]. The structural diversity of polyphenols provides a broad chemical space for the discovery of novel diet-based DPP-4 inhibitors.

Recent advances in computational approaches have significantly accelerated the drug discovery process. In silico methods, particularly machine learning (ML) and molecular docking, offer efficient and cost-effective strategies for screening large compound libraries prior to experimental validation. ML-based cheminformatics approaches enable the prediction of biological activity using molecular descriptors and fingerprints with competitive accuracy [9,10]. Meanwhile, molecular docking provides mechanistic insights into ligand–receptor interactions at the atomic level, including binding modes and key amino acid residues involved in inhibition [11]. The integration of ML and docking is increasingly recognized as a powerful approach for identifying bioactive compounds from natural sources.

Despite the recognized potential of dietary polyphenols and the robustness of computational methods, large-scale screening studies that integrate the predictive power of machine learning with the structural accuracy of molecular docking, specifically applied to comprehensive food polyphenol databases, remain limited. Therefore, this study aims to develop an integrated in silico pipeline to identify dietary polyphenols as potential DPP-4 inhibitors. The machine learning model was trained using bioactivity data from the ChEMBL database, while virtual screening was conducted on a polyphenol library derived from the Phenol-Explorer database. This approach is expected to efficiently prioritize candidate compounds with potential antidiabetic activity and contribute to the development of functional food-based strategies for T2DM management.

## 2. EXPERIMENTAL

### *2.1. Preparation of the DPP-4 Bioactivity Dataset*

Bioactivity data of ligands targeting the dipeptidyl peptidase-4 (DPP-4) enzyme were retrieved from the ChEMBL database (Target ID: CHEMBL284) using the ChEMBL API, with a focus on the half-maximal inhibitory concentration ($IC_{50}$) parameter. Compounds were categorized into three classes: active ($IC_{50} \leq 1{,}000$ nM), intermediate ($1{,}000 < IC_{50} \leq 10{,}000$ nM), and inactive ($IC_{50} > 10{,}000$ nM). Compounds with intermediate activity were excluded to reduce biological ambiguity and improve model robustness [12]. The final dataset consisted of a binary classification comprising 4,248 compounds, with class imbalance (ratio 3.4:1), including 3,288 active compounds and 960 inactive compounds.


e-ISSN : 2622-4968, p-ISSN : 2622-1349 

### 2.2. Preparation of the Screening Library from Phenol-Explorer

The screening library was obtained from the Phenol-Explorer database version 3.6 [13,14], which provides metadata for 501 phenolic compounds along with quantitative composition data across 458 food items. From this database, 492 entries containing chemical structure representations (SMILES) were available and subsequently merged with the metadata of 501 compounds based on unique identifiers prior to structural standardization. Molecular standardization was performed using RDKit [15,16] through a sequential workflow, including metal disconnection, largest fragment selection to remove counter-ions, charge normalization, and molecular neutralization, followed by conversion into canonical SMILES format.

Subsequently, drug-likeness filtering was applied using Extended Lipinski criteria (molecular weight < 500 Da, LogP between −3 and 5, hydrogen bond donors ≤ 5, hydrogen bond acceptors ≤ 10), Veber rules (topological polar surface area ≤ 140 Å², rotatable bonds ≤ 10), and a quantitative estimate of drug-likeness (QED) threshold of ≥ 0.3 [17,18]. To reduce redundancy arising from high structural similarity, clustering was performed based on the Tanimoto similarity coefficient (Tc > 0.85) using Morgan fingerprints (radius 2, 1024-bit) [19]. From each cluster, a representative compound was selected based on the highest occurrence frequency in food sources (n_foods), resulting in a final library of 162 unique compounds for subsequent virtual screening.

### 2.3. Molecular Featurization for Machine Learning

Each molecule was represented as a vector of 2,053 features, consisting of: (1) Morgan fingerprints (ECFP4, radius = 2, 2048-bit) capturing local substructural information, and (2) five RDKit physicochemical descriptors (MolWt, MolLogP, TPSA, NumHBD, NumHBA) representing global molecular properties [20]. A variance threshold filter (threshold = 0.05) was applied to remove non-informative features. The dataset was then partitioned using a scaffold-based splitting strategy based on Murcko scaffolds into training (70%), validation (15%), and test (15%) sets, ensuring no scaffold overlap between subsets [21]. The same featurization pipeline, variance filtering, and StandardScaler were consistently applied to the Phenol-Explorer library.

### 2.4. Machine Learning Model Training and Evaluation

Six machine learning algorithms were evaluated: Random Forest (RF), XGBoost, Support Vector Machine (SVM), Logistic Regression, K-Nearest Neighbors (KNN), and Gradient Boosting. Class imbalance was addressed using the Synthetic Minority Over-sampling Technique (SMOTE), applied exclusively to the training set, resulting in an active ratio of 1:2. The StandardScaler was fitted solely on the SMOTE-augmented training set [22,23]. Hyperparameter optimization was conducted using randomized search with 30 iterations and stratified 5-fold cross-validation, employing the Matthews Correlation Coefficient (MCC) as the primary evaluation metric [24]. The optimal prediction threshold for each model was determined from the Precision–Recall curve based on maximum F1-score. An ensemble model was constructed using a soft-voting strategy, with an optimized ensemble threshold of 0.527.


e-ISSN : 2622-4968, p-ISSN : 2622-1349 

### *2.5. Virtual Screening*

Virtual screening was performed by applying the trained ensemble model to the Phenol-Explorer library, which had been featurized using the same pipeline as in the training phase (including identical variance filtering and StandardScaler). Tree-based models (RF, XGBoost, and Gradient Boosting) utilized unscaled features, whereas SVM, Logistic Regression, and KNN were applied to scaled features. Compounds with an ensemble probability ≥ 0.527 were classified as potential active DPP-4 inhibitors and subsequently mapped to their corresponding dietary sources using the Phenol-Explorer composition data.

### *2.6. Receptor Preparation and Molecular Docking*

The crystal structure of dipeptidyl peptidase-4 (DPP-4) (PDB ID: 4A5S; resolution 2.3 Å) was prepared by removing water molecules, co-crystallized ligands, and ions, followed by the addition of hydrogen atoms and Gasteiger charges using AutoDockTools. Molecular docking was performed using AutoDock Vina with a grid center at (13.782, 35.764, 55.517 Å), grid dimensions of 40 × 40 × 40 Å, exhaustiveness set to 32, num_modes to 9, and an energy range of 3 kcal/mol.

Protocol validation was conducted through redocking of the reference ligand sitagliptin using an MCS-based RMSD approach implemented in RDKit, yielding an RMSD value of 1.966 Å (< 2.0 Å), indicating a valid docking protocol. A total of 11 compounds were prepared from SMILES using Open Babel (gen3D, pH 7.4, Gasteiger charges, MMFF94 minimization) [25,26]. Interaction analysis was focused on the top candidates (hesperetin, pelargonidin, daidzein, and naringenin) using Discovery Studio Visualizer to identify key ligand–receptor interactions.

## 3. RESULTS AND DISCUSSION

### *3.1. Dataset Characteristics and Machine Learning Model Performance*

The DPP-4 bioactivity dataset derived from ChEMBL comprised 4,248 binary-class compounds (Table 1), with an inactive ratio of 3.4:1 following the removal of intermediate-class compounds. This significant class imbalance necessitated specific handling to prevent the model from being biased toward predicting the majority class (inactive). The application of SMOTE to the training set adjusted the class ratio to 1:2 (active), thereby providing a more balanced learning signal without altering the distribution of the validation and test sets.

The comparative evaluation results of six individual machine learning (ML) algorithms and one ensemble model on the test set are presented in Table 1 and Figure 1. The soft-voting ensemble approach, which integrates all six algorithms, achieved the highest overall predictive performance, with AUC-ROC = 0.9739, AUC-PR = 0.9919, and MCC = 0.8334—matching the best-performing individual model, XGBoost. The high MCC value indicates that the model maintains balanced predictive performance across both active (positive) and inactive (negative) classes, without bias toward either class—a critical characteristic for imbalanced bioactivity datasets [27].


e-ISSN : 2622-4968, p-ISSN : 2622-1349 

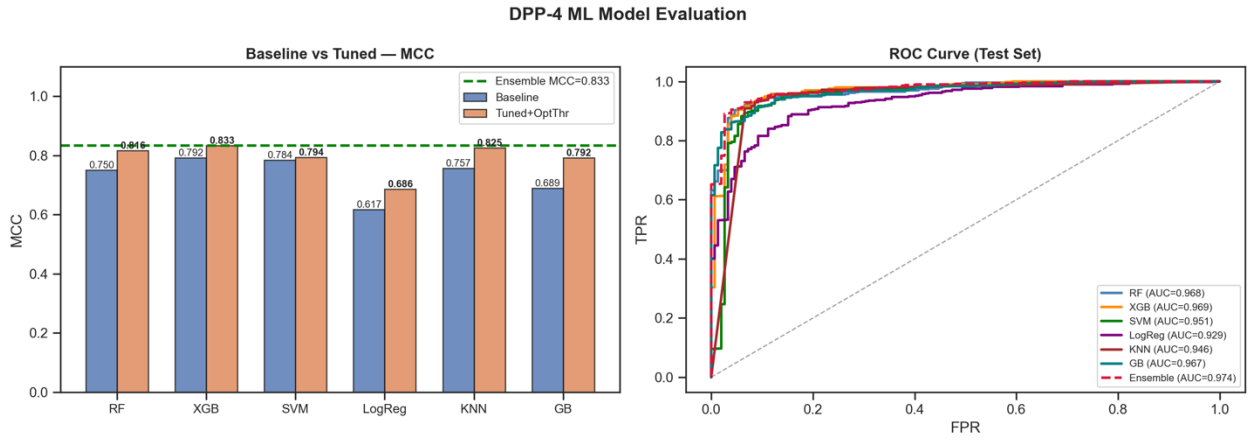



**Figure 1.** Performance evaluation of machine learning models for DPP-4 prediction. (Left) Comparison of MCC values between baseline and tuned models with optimal thresholds across six algorithms; the dashed green line indicates the ensemble MCC of 0.833. (Right) ROC curves on the test set; ensemble AUC-ROC = 0.974.

**Table 1.** Comparison of machine learning model performance on the test set

| Model | Threshold | MCC | AUC-ROC | AUC-PR | F1-score |
|---|---|---|---|---|---|
| Random Forest | 0,579 | 0,8157 | 0,9682 | 0,9902 | 0,9550 |
| XGBoost | 0,570 | 0,8334 | 0,9692 | 0,9893 | 0,9604 |
| SVM | 0,576 | 0,7936 | 0,9505 | 0,9756 | 0,9530 |
| Logistic Regression | 0,418 | 0,6860 | 0,9291 | 0,9769 | 0,9232 |
| KNN | 0,358 | 0,8249 | 0,9461 | 0,9719 | 0,9584 |
| Gradient Boosting | 0,558 | 0,7918 | 0,9673 | 0,9901 | 0,9502 |
| Ensemble (6 model) | 0,527 | 0,8334 | 0,9739 | 0,9919 | 0,9604 |

*Thresholds were optimized using Precision–Recall curves based on maximum F1-score. MCC = Matthews Correlation Coefficient; AUC-PR = Area Under the Precision–Recall Curve.*

Among individual models, XGBoost and KNN demonstrated superior performance, with MCC values of 0.8334 and 0.8249, respectively. The strength of XGBoost lies in its gradient boosting mechanism based on decision trees, which iteratively minimizes the loss function by reweighting residual errors from previous


e-ISSN : 2622-4968, p-ISSN : 2622-1349

predictions [28]. KNN, on the other hand, operates by identifying local structural similarities among ligands in a multidimensional feature space without requiring assumptions about data distribution, allowing it to effectively capture local non-linear patterns within the DPP-4 chemical space.

In contrast, Logistic Regression exhibited the lowest performance, with an MCC of 0.6860, confirming that the structure–activity relationship of DPP-4 inhibitors is inherently non-linear and cannot be optimally captured by linear models—consistent with findings from other cheminformatics-based bioactivity modeling studies [29,30]. Although all models achieved AUC-ROC values above 0.92, this metric may present an overly optimistic view in virtual screening scenarios dominated by inactive compounds. The AUC-PR metric provides a more reliable evaluation under class imbalance conditions [31], and the ensemble AUC-PR value of 0.9919 demonstrates the model's ability to minimize false positives without sacrificing sensitivity in detecting active compounds—a critical requirement for reducing experimental failure costs in downstream wet-lab validation [32].

The consistent improvement in model performance can be attributed to two key strategies. First, the application of SMOTE (Synthetic Minority Over-sampling Technique) mitigates model bias toward the majority class by generating synthetic active samples along the feature space [33,34]. Second, instead of using the conventional decision threshold of 0.5, the prediction threshold was dynamically optimized using the maximum F1-score from the Precision–Recall curve on the test set, resulting in an optimal ensemble threshold of 0.527, which effectively balances precision and recall.

### *3.2. Virtual Screening Results from the Phenol-Explorer Database*

Computational intelligence–based virtual screening represents an important approach for accelerating natural product–based drug discovery. In this study, an ensemble model using an optimized probability threshold of 0.527 was applied to 162 compounds from the Phenol-Explorer database, resulting in 11 predicted active candidates (Table 2).

**Table 2.** Active DPP-4 Candidates Identified from Phenol-Explorer Virtual Screening

| Compounds | Class/Subclass | Prob-DPP4 | MW (Da) | LogP | QED | Food Group Utama |
|---|---|---|---|---|---|---|
| Juglone | Naphtoquinone | 0,6404 | 174,2 | 1,33 | 0,645 | Alcoholic beverages, Seeds |
| Genistein | Isoflavonoid | 0,6386 | 270,2 | 2,58 | 0,632 | Alcoholic beverages, Non-alcoholic beverages, Seasonings, Seeds |
| Pelargonidin | Anthocyanin | 0,6336 | 271,2 | 3,20 | 0,510 | Fruits and fruit products, Seeds |
| 1,4-Naphthoquinone | Naphtoquinone | 0,6032 | 158,2 | 1,62 | 0,575 | Alcoholic beverages |
| Rosmadial | Phenolic terpene | 0,5999 | 344,4 | 3,51 | 0,515 | Seasonings |
| Daidzein | Isoflavonoid | 0,5884 | 254,2 | 2,87 | 0,700 | Alcoholic beverages, Non-alcoholic beverages, Seasonings, Seeds |


e-ISSN : 2622-4968, p-ISSN : 2622-1349

| Compounds | Class/Subclass | Prob-DPP4 | MW (Da) | LogP | QED | Food Group Utama |
|---|---|---|---|---|---|---|
| Naringenin | Flavanone | 0,5851 | 272,3 | 2,51 | 0,742 | Alcoholic beverages, Fruits, Non-alcoholic beverages, Seasonings, Seeds, Vegetables |
| Resveratrol | Stilbenoid | 0,5614 | 228,2 | 2,97 | 0,692 | Alcoholic beverages, Coffee and cocoa, Fruits, Non-alcoholic beverages, Oils, Seasonings, Seeds |
| Mellein | Hydroxycoumarin | 0,5409 | 178,2 | 1,49 | 0,612 | Coffee and cocoa |
| Hesperetin | Flavanone | 0,5403 | 302,3 | 2,52 | 0,789 | Alcoholic beverages, Fruits, Non-alcoholic beverages, Seasonings, Vegetables |
| Piceatannol | Stilbenoid | 0,5401 | 244,2 | 2,68 | 0,483 | Alcoholic beverages, Fruits and fruit products |

*Prob-DPP4 = ensemble prediction probability (threshold = 0.527); QED = Quantitative Estimate of Drug-likeness; MW = molecular weight (Da). Compounds are ranked in descending order of predicted probability..*

Structurally, the predicted active compounds were predominantly flavonoids, followed by other polyphenols and stilbenes. From a medicinal chemistry perspective, several compounds such as hesperetin, naringenin, and daidzein exhibited high Quantitative Estimate of Drug-likeness (QED) values (> 0.7), indicating favorable physicochemical properties for oral drug candidacy [17]. In contrast, piceatannol showed a relatively lower QED value, likely due to its higher number of hydroxyl groups, which may reduce membrane permeability despite its potential binding affinity toward the target protein.

Other physicochemical parameters, including LogP and molecular weight (MW), further supported the drug-likeness of the identified candidates. Most compounds fell within the optimal LogP range (1–3), contributing to a balanced profile between aqueous solubility and membrane permeability [35]. Additionally, low molecular weight compounds such as juglone and 1,4-naphthoquinone offer greater structural flexibility for further optimization in lead development processes.

### *3.3. Molecular Docking Analysis*

Redocking of the co-crystallized ligand into the active site of DPP-4 (PDB ID: 4A5S) yielded a binding energy of ΔG = −9.905 kcal/mol, while sitagliptin exhibited ΔG = −9.640 kcal/mol and was used as a reference compound. All 11 candidates obtained from virtual screening were subsequently docked against the DPP-4 receptor (4A5S). The remaining seven candidates (rosmadial, genistein, piceatannol, resveratrol, juglone, 1,4-naphthoquinone, and mellein) showed binding energies ranging from −7.050 to −8.259 kcal/mol. Detailed discussion was focused on four compounds with the strongest binding affinities (ΔG < −8.3 kcal/mol), namely hesperetin, pelargonidin, daidzein, and naringenin (Table 3).


e-ISSN : 2622-4968, p-ISSN : 2622-1349 

**Table 3.** Molecular Docking Results of Active DPP-4 Candidates against the 4A5S Receptor

| Compounds | ΔG (kcal/mol) | vs Sitagliptin | Key Interactions |
|---|---|---|---|
| Hesperetin | −8,548 | −1,092 | Tyr631, Ser209 (H-bond); Pi-Anion Glu206; Pi-Pi Tyr662, Trp659, Tyr666; Alkyl Val656, Phe357 |
| Pelargonidin | −8,432 | −1,208 | Ser209, Arg125 (H-bond); Pi-Anion Glu205, Glu206, Phe357 |
| Daidzein | −8,386 | −1,254 | Tyr547, Tyr662, Arg358 (H-bond); Pi-Pi Stacked; tanpa interaksi unfavorable |
| Naringenin | −8,321 | −1,319 | Ser209, Phe357 (H-bond); Pi-Anion Glu206; unfavorable: Arg669, Glu206 |

*ΔG = binding free energy (kcal/mol); more negative values indicate higher affinity. vs. sitagliptin = ΔG difference compared to the reference (−9.640 kcal/mol).*

Hesperetin exhibited the most favorable binding energy among all candidates (ΔG = −8.548 kcal/mol), followed by pelargonidin (−8.432 kcal/mol), daidzein (−8.386 kcal/mol), and naringenin (−8.321 kcal/mol). All four compounds demonstrated ΔG values below −8.0 kcal/mol, indicating strong affinity toward the DPP-4 active site. Although these values do not surpass that of sitagliptin, they are considered to possess significant biological potential as enzyme inhibitors [36,37].

Based on molecular docking simulations and 2D/3D interaction visualizations, detailed insights into the inhibition mechanism of DPP-4 by the four top flavonoid candidates were obtained (Figure 2). The DPP-4 active site consists of several subsites: the S1 subsite (hydrophobic, enriched with aromatic residues such as Tyr662, Tyr666, and Trp659), the S2 subsite (hydrophilic/charged, involving Glu205, Glu206, and Arg125), and the catalytic triad (Ser630, Asp708, and His740) [38,39].


e-ISSN : 2622-4968, p-ISSN : 2622-1349

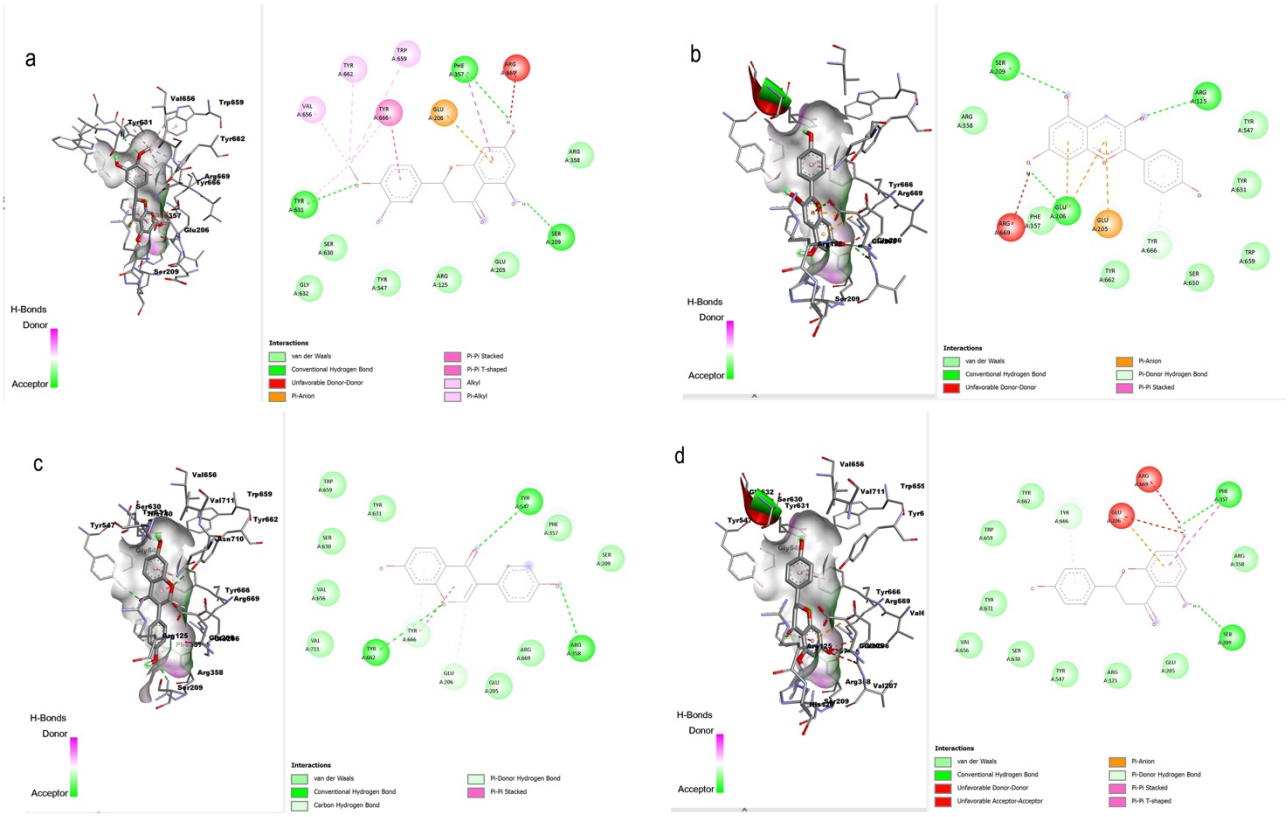



**Figure 2.** 3D and 2D visualization of molecular docking of DPP-4 (PDB: 4A5S) with (a) hesperetin, (b) pelargonidin, (c) daidzein, and (d) naringenin.

Hesperetin demonstrated the most favorable binding profile (ΔG = −8.548 kcal/mol; Figure 2a), which can be attributed to its diverse interaction network within the active site. Conventional hydrogen bonds were observed with Tyr631 and Ser209, where interaction with Ser209 plays a crucial role in stabilizing ligand orientation within the S2 subsite [40]. Additionally, electrostatic contributions were evident through π–anion interactions with Glu206. Various hydrophobic and aromatic interactions, including π–π stacking, π–π T-shaped, alkyl, and π–alkyl interactions, were identified with residues Tyr662, Trp659, Tyr666, Phe357, and Val656 in the S1 subsite. These interactions likely contribute significantly to ligand–protein complex stability. The presence of a methoxy (–$OCH_3$) group on the B ring of hesperetin is suggested to enhance steric flexibility, enabling deeper accommodation within the hydrophobic pocket. Although an unfavorable donor–donor interaction with Arg669 was observed, its destabilizing effect is likely compensated by surrounding hydrophobic interactions and van der Waals forces.

In contrast to hesperetin, pelargonidin—an anthocyanin commonly found in strawberries and black beans—also exhibited strong binding affinity (ΔG = −8.432 kcal/mol; Figure 2b). This affinity is likely influenced by its flavylium cation character, which enhances electrostatic interactions within the active site [41]. Conventional hydrogen bonds were observed with Ser209 and Arg125, where Arg125 acts as a key residue in the S2 subsite for ligand binding and orientation [42]. In addition, π–anion interactions with Glu205 and Glu206—negatively charged residues within the S2 subsite—likely contribute significantly to the stability of


e-ISSN : 2622-4968, p-ISSN : 2622-1349

the ligand–protein complex [43]. Additional van der Waals contacts with surrounding residues further support ligand stabilization within the binding pocket [44]. Although an unfavorable donor–donor interaction with Arg669 was observed, its destabilizing effect is likely mitigated by the dominance of hydrogen bonding and electrostatic interactions. Overall, this interaction profile indicates that pelargonidin effectively engages key residues within the S2 subsite of DPP-4.

Daidzein, an isoflavone with the B ring positioned at C3 and abundant in soy products and legumes (Figure 2c), ranked third with ΔG = −8.386 kcal/mol. This structural difference significantly alters its interaction profile compared to conventional flavones. Hydrogen bonds were formed with Tyr547, Tyr662, and Arg358 via its carbonyl group. π–π stacking interactions were observed with aromatic residues in the S1 subsite. Notably, daidzein exhibited no unfavorable interactions, either donor–donor or acceptor–acceptor. The absence of steric or electrostatic repulsion likely contributes to its highly stable geometric orientation within the binding pocket, despite having slightly higher ΔG compared to pelargonidin [41].

Naringenin, which shares the same flavanone backbone as hesperetin but lacks the methoxy (–$OCH_3$) group on the B ring (Figure 2d), showed the weakest binding among the four candidates (ΔG = −8.321 kcal/mol). Hydrogen bonds were observed with Ser209 and Phe357, while π–anion interaction with Glu206 contributed to electrostatic stabilization. Limited aromatic interactions, such as π–π T-shaped interaction with Phe357, were also identified, whereas other hydrophobic interactions were relatively minimal. Notably, two unfavorable interactions were detected: donor–donor interaction with Arg669 and acceptor–acceptor interaction with Glu206. These unfavorable interactions likely contribute to its reduced binding affinity. The difference in affinity between hesperetin and naringenin may be attributed to the absence of the methoxy group in naringenin, which reduces its ability to form additional hydrophobic interactions and optimally adapt within the binding pocket (Ayaz et al., 2026).

Comparative analysis of the interaction profiles of these four flavonoids provides key insights into structure–activity relationships (SAR) for DPP-4 inhibition. First, residues Ser209 and Glu205/206 appear to play critical roles, as high-affinity ligands (hesperetin and pelargonidin) consistently interact strongly with these residues, which are responsible for binding the N-terminal region of natural DPP-4 substrates [47,48]. Second, substitutions on the B ring significantly influence binding affinity: the presence of a methoxy group in hesperetin enhances hydrophobic volume, enabling additional alkyl and π–alkyl interactions with Val656 and Tyr662, which are absent in naringenin (containing only hydroxyl groups), explaining the ~0.23 kcal/mol difference in binding energy. Third, unfavorable interactions impact complex stability: a single unfavorable interaction in hesperetin and pelargonidin can be compensated by strong interaction networks, whereas the accumulation of two unfavorable interactions in naringenin directly reduces stability. Daidzein, which lacks unfavorable interactions entirely, exhibits the most stable geometric binding orientation despite having slightly higher ΔG compared to pelargonidin.

### *3.4. Food Sources of Active Candidates and Nutritional Implications*

The developed in silico pipeline successfully addressed the primary objective of this study by systematically identifying dietary polyphenols as potential DPP-4 inhibitors from the comprehensive Phenol-Explorer library. From a nutritional perspective, the distribution of food sources associated with the identified active candidates (Table 4) is highly relevant for dietary recommendations.


e-ISSN : 2622-4968, p-ISSN : 2622-1349 

Naringenin (present in 58 food items, including citrus fruits, berries, cruciferous vegetables, tea, and grapes) and hesperetin (found in 52 food items, including citrus fruits, black tea, and vegetables) exhibited the broadest food accessibility, making them high-priority candidates for future nutritional intervention studies. Daidzein and genistein are primarily derived from soy-based products, a food group whose consumption can be readily increased in dietary patterns. Pelargonidin is abundant in strawberries and black beans, whereas resveratrol is commonly found in red grapes and berries.

This distribution supports the recommendation of polyphenol-rich, plant-based dietary patterns as a complementary strategy in T2DM management [49,50]. However, the in vivo bioavailability of polyphenols is influenced by multiple factors, including the food matrix, processing methods, gut metabolism, and individual genetic variability—factors that cannot be fully captured by in silico approaches. Therefore, experimental validation through in vitro assays (e.g., direct DPP-4 inhibition), bioavailability studies, and controlled clinical trials remains essential to confirm the clinical relevance of these findings.

Among the identified compounds, hesperetin and naringenin (derived from citrus fruits and vegetables), as well as daidzein (from soy-based foods), are prioritized for further investigation due to their favorable docking energies, stable interaction profiles, and wide dietary availability.

## 4. CONCLUSION

An integrated in silico pipeline combining machine learning and molecular docking successfully identified dietary polyphenols as potential DPP-4 inhibitors. The soft-voting ensemble demonstrated excellent predictive performance (MCC = 0.8334; AUC-ROC = 0.9739), with reliable docking validation (RMSD = 1.966 Å). Screening of 162 compounds yielded 11 active candidates, with hesperetin ($\Delta G = -8.548$ kcal/mol) emerging as the most promising, followed by pelargonidin, daidzein, and naringenin. These compounds, naturally abundant in citrus fruits, soy products, berries, and spices, represent promising natural antidiabetic agents. Further in vitro, bioavailability, and clinical validation remains warranted.

## ACKNOWLEDGEMENT

Not applicable

## REFERENCES

1. Ji, Q., Chai, S., Zhang, R., Li, J., Zheng, Y., & Rajpathak, S. (2024). Prevalence and co-prevalence of comorbidities among Chinese adult patients with type 2 diabetes mellitus: a cross-sectional, multicenter, retrospective, observational study based on 3B study database. *Frontiers in Endocrinology*, *15*. https://doi.org/10.3389/fendo.2024.1362433
2. Gheonea, T. C., Șurlin, P., Nicolae, F. M., Gheorghe, D. N., Popescu, D. M., & Rogoveanu, I. (2024). Dipeptidyl-Peptidase-4 and Glucagon-like-Peptide-1, a Link in the Connection between Periodontitis and Diabetes Mellitus—What Do We Know So Far?—A Scoping Review. *Journal of Clinical Medicine*, *13*(3), 903. https://doi.org/10.3390/jcm13030903
3. Alrouji, M., Al-kuraishy, H. M., Al-buhadily, A. K., Al-Gareeb, A. I., Elekhnawy, E., & Batiha, G. E.-S. (2023). DPP-4 inhibitors and type 2 diabetes mellitus in Parkinson's disease: a mutual relationship. *Pharmacological Reports*, *75*(4), 923–936. https://doi.org/10.1007/s43440-023-00500-5
4. Zhang, J., Xu, S., Liu, X., Zhang, J., Hu, S., Liu, X., Yang, C., & Fang, Y. (2024). Time trends and regional variation in utilization of antidiabetic medicines in China, 2015–2022. *Diabetes, Obesity and Metabolism*, *26*(7), 2752–2760. https://doi.org/10.1111/dom.15594



e-ISSN : 2622-4968, p-ISSN : 2622-1349 

5. Karagiannis, T., Bekiari, E., & Tsapas, A. (2023). Socioeconomic aspects of incretin-based therapy. *Diabetologia*, *66*(10), 1859–1868. https://doi.org/10.1007/s00125-023-05962-z
6. Aryal, D., Joshi, S., Thapa, N. K., Chaudhary, P., Basaula, S., Joshi, U., Bhandari, D., Rogers, H. M., Bhattarai, S., Sharma, K. R., Regmi, B. P., & Parajuli, N. (2024). Dietary phenolic compounds as promising therapeutic agents for diabetes and its complications: A comprehensive review. *Food Science & Nutrition*, *12*(5), 3025–3045. https://doi.org/10.1002/fsn3.3983
7. Ansari, P., Samia, J. F., Khan, J. T., Rafi, M. R., Rahman, Md. S., Rahman, A. B., Abdel-Wahab, Y. H. A., & Seidel, V. (2023). Protective Effects of Medicinal Plant-Based Foods against Diabetes: A Review on Pharmacology, Phytochemistry, and Molecular Mechanisms. *Nutrients*, *15*(14), 3266. https://doi.org/10.3390/nu15143266
8. Nazir, M., Haq, M. A., Ali, S. A., Ghufran Saeed, S. M., Ajmal, M. A., Nisar, M., Khan, T. A., Tsoupras, A., & Naz, S. (2025). Antioxidants and Antidiabetic Potential of Polyphenolic Fractions and Crude Extracts of *Rhus typhina* Fruit, *Punica granatum* L. Peel, and *Terminalia catappa* L. Leaves: In Vitro and In Vivo Evaluation. *Chemistry & Biodiversity*, *22*(7). https://doi.org/10.1002/cbdv.202500020
9. D'Anna, L., Contino, S., Marinello, R., Fares, J., De Simone, G., Monari, A., Barbault, F., Barone, G., Terenzi, A., & Perricone, U. (2026). AGAPE (Computational G-Quadruplex Stabilization Prediction): The First Machine Learning Workflow for G-Quadruplex Stabilization Prediction. *ACS Omega*, *11*(21), 31744–31756. https://doi.org/10.1021/acsomega.6c03072
10. Gadiya, Y., Genilloud, O., Bilitewski, U., Brönstrup, M., von Berlin, L., Attwood, M., Gribbon, P., & Zaliani, A. (2025). Predicting Antimicrobial Class Specificity of Small Molecules Using Machine Learning. *Journal of Chemical Information and Modeling*, *65*(5), 2416–2431. https://doi.org/10.1021/acs.jcim.4c02347
11. Samanta, P. N., Majumdar, D., & Leszczynski, J. (2023). Elucidating Atomistic Insight into the Dynamical Responses of the SARS-CoV-2 Main Protease for the Binding of Remdesivir Analogues: Leveraging Molecular Mechanics To Decode the Inhibition Mechanism. *Journal of Chemical Information and Modeling*, *63*(11), 3404–3422. https://doi.org/10.1021/acs.jcim.3c00105
12. Seal, S., Mahale, M., García-Ortegón, M., Joshi, C. K., Hosseini-Gerami, L., Beatson, A., Greenig, M., Shekhar, M., Patra, A., Weis, C., Mehrjou, A., Badré, A., Paisley, B., Lowe, R., Singh, S., Shah, F., Johannesson, B., Williams, D., Rouquie, D., … Bender, A. (2025). Machine Learning for Toxicity Prediction Using Chemical Structures: Pillars for Success in the Real World. *Chemical Research in Toxicology*, *38*(5), 759–807. https://doi.org/10.1021/acs.chemrestox.5c00033
13. Rothwell, J. A., Perez-Jimenez, J., Neveu, V., Medina-Remon, A., M'Hiri, N., Garcia-Lobato, P., Manach, C., Knox, C., Eisner, R., Wishart, D. S., & Scalbert, A. (2013). Phenol-Explorer 3.0: a major update of the Phenol-Explorer database to incorporate data on the effects of food processing on polyphenol content. *Database*, *2013*(0), bat070–bat070. https://doi.org/10.1093/database/bat070
14. Neveu, V., Perez-Jimenez, J., Vos, F., Crespy, V., du Chaffaut, L., Mennen, L., Knox, C., Eisner, R., Cruz, J., Wishart, D., & Scalbert, A. (2010). Phenol-Explorer: an online comprehensive database on polyphenol contents in foods. *Database*, *2010*(0), bap024–bap024. https://doi.org/10.1093/database/bap024
15. Müller, S. (2025). How to crack a SMILES: automatic crosschecked chemical structure resolution across multiple services using MoleculeResolver. *Journal of Cheminformatics*, *17*(1), 117. https://doi.org/10.1186/s13321-025-01064-7
16. Dolciami, D., Villasclaras-Fernandez, E., Kannas, C., Meniconi, M., Al-Lazikani, B., & Antolin, A. A. (2022). canSAR chemistry registration and standardization pipeline. *Journal of Cheminformatics*, *14*(1), 28. https://doi.org/10.1186/s13321-022-00606-7
17. Montenegro-Herrera, S. A., Sosa, A., Echeverri-Jiménez, I., Castaño-Valencia, R. S., & Jerez-Valderrama, A. M. (2025). Machine Learning-Based QSAR Screening of Colombian Medicinal Flora for Potential Antiviral Compounds Against Dengue Virus: An In Silico Drug Discovery Approach. *Pharmaceuticals*, *18*(12), 1906. https://doi.org/10.3390/ph18121906
18. Bickerton, G. R., Paolini, G. V., Besnard, J., Muresan, S., & Hopkins, A. L. (2012). Quantifying the chemical beauty of drugs. *Nature Chemistry*, *4*(2), 90–98. https://doi.org/10.1038/nchem.1243
19. Shanbhogue, R. V., Gandhi, N. S., B., S. P., & Navada, S. G. (2026). Machine learning based approaches for structure activity relationship analysis of heparanase inhibitors. *Scientific Reports*. https://doi.org/10.1038/s41598-026-55098-4
20. Liu, K., Cui, H., Yu, X., Li, W., & Han, W. (2025). Predicting cardiotoxicity in drug development: A deep learning approach. *Journal of Pharmaceutical Analysis*, *15*(8), 101263. https://doi.org/10.1016/j.jpha.2025.101263




e-ISSN : 2622-4968, p-ISSN : 2622-1349 

21. Xia, X., Zhu, C., Zhong, F., & Liu, L. (2024). TransCDR: a deep learning model for enhancing the generalizability of drug activity prediction through transfer learning and multimodal data fusion. *BMC Biology*, *22*(1), 227. https://doi.org/10.1186/s12915-024-02023-8
22. Han, S., & Jung, H. (2024). NATE: Non-pArameTric approach for Explainable credit scoring on imbalanced class. *PLOS ONE*, *19*(12), e0316454. https://doi.org/10.1371/journal.pone.0316454
23. Al-Ahmari, S., & Nadeem, F. (2025). Improving Surgical Site Infection Prediction Using Machine Learning: Addressing Challenges of Highly Imbalanced Data. *Diagnostics*, *15*(4), 501. https://doi.org/10.3390/diagnostics15040501
24. Almadhoun, M. B., & Burhanuddin, M. (2025). Optimizing Feature Selection and Machine Learning Algorithms for Early Detection of Prediabetes Risk: Comparative Study. *JMIR Bioinformatics and Biotechnology*, *6*, e70621–e70621. https://doi.org/10.2196/70621
25. Burns, J. W., & Rogers, D. M. (2023). QuantumScents: Quantum-Mechanical Properties for 3.5k Olfactory Molecules. *Journal of Chemical Information and Modeling*, *63*(23), 7330–7337. https://doi.org/10.1021/acs.jcim.3c01338
26. Vitalis, A., Parker, D., Radler, F., Vargas-Rosales, P. A., Zhang, Y., Marchand, J., & Caflisch, A. (2025). Parallelized Tools for the Preparation and Curation of Large Libraries for Virtual Screening. *Journal of Computational Chemistry*, *46*(7). https://doi.org/10.1002/jcc.70074
27. Astigarraga, M., Sánchez-Ruiz, A., & Colmenarejo, G. (2025). *Conformal Prediction-based Machine Learning in Cheminformatics: Current Applications and New Challenges*. https://doi.org/10.26434/chemrxiv-2025-p36vt
28. Noorunnahar, M., Chowdhury, A. H., & Mila, F. A. (2023). A tree based eXtreme Gradient Boosting (XGBoost) machine learning model to forecast the annual rice production in Bangladesh. *PLOS ONE*, *18*(3), e0283452. https://doi.org/10.1371/journal.pone.0283452
29. Acosta-Murillo, R., Ortiz-Bayliss, J. C., & Zapata-Morin, P. A. (2026). Enhancing cancer drug discovery: QSAR modeling with machine learning and chemical representations. *PLOS One*, *21*(3), e0343654. https://doi.org/10.1371/journal.pone.0343654
30. Yang, S., Li, S., & Chang, J. (2022). Discovery of Cobimetinib as a novel A-FABP inhibitor using machine learning and molecular docking-based virtual screening. *RSC Advances*, *12*(21), 13500–13510. https://doi.org/10.1039/D2RA01057G
31. Chirila, C.-B., & Crisan, L. (2026). Harnessing machine learning models to repurpose drugs targeting HIV-1 integrase, protease, and reverse transcriptase. *Scientific Reports*, *16*(1), 18146. https://doi.org/10.1038/s41598-026-48926-0
32. Bagdad, Y., Sisquellas, M., Arthur, M., & Miteva, M. A. (2024). Machine Learning and Deep Learning Models for Predicting Noncovalent Inhibitors of AmpC β-Lactamase. *ACS Omega*, *9*(40), 41334–41342. https://doi.org/10.1021/acsomega.4c03834
33. Goldwaser, E., Laurent, C., Lagarde, N., Fabrega, S., Nay, L., Villoutreix, B. O., Jelsch, C., Nicot, A. B., Loriot, M.-A., & Miteva, M. A. (2022). Machine learning-driven identification of drugs inhibiting cytochrome P450 2C9. *PLOS Computational Biology*, *18*(1), e1009820. https://doi.org/10.1371/journal.pcbi.1009820
34. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953
35. Rai, M., Singh, A. V., Paudel, N., Kanase, A., Falletta, E., Kerkar, P., Heyda, J., Barghash, R. F., Pratap Singh, S., & Soos, M. (2023). Herbal concoction Unveiled: A computational analysis of phytochemicals' pharmacokinetic and toxicological profiles using novel approach methodologies (NAMs). *Current Research in Toxicology*, *5*, 100118. https://doi.org/10.1016/j.crtox.2023.100118
36. Abdelrahman, A. H. M., Mekhemer, G. A. H., Sidhom, P. A., El-Tayeb, M. A., Khan, S., & Ibrahim, M. A. A. (2025). Exploration of African natural products as VP35 inhibitors to combat Marburg virus infection: Molecular docking, molecular dynamics, and quantum mechanical computations. *PLOS One*, *20*(10), e0334160. https://doi.org/10.1371/journal.pone.0334160
37. Akinnusi, P. A., Olubode, S. O., Alade, A. A., Ashimi, A. A., Onawola, O. L., Agbolade, A. O., Emeka, A. P., Shodehinde, S. A., & Adeniran, O. Y. (2023). Potential Inhibitory Biomolecular Interactions of Natural Compounds With Different Molecular Targets of Diabetes. *Bioinformatics and Biology Insights*, *17*. https://doi.org/10.1177/11779322231167970
38. Tsahnang Fofack, H. M., Mbah Bake, M., Petry, S., Ateba, B. A., Amoa Onguéné, P., Mohammad-Salim, H., Ntie-Kang, F., Mbaze, L. M., Vakal, S., & Kenfack, C. A. (2024). Identification of potential dipeptidyl peptidase IV




e-ISSN : 2622-4968, p-ISSN : 2622-1349 

inhibitors from the ConMedNP library by virtual screening, and molecular dynamics methods. *Heliyon*, *10*(15), e35191. https://doi.org/10.1016/j.heliyon.2024.e35191

39. Antony, P., Baby, B., Aleissaee, H. M., & Vijayan, R. (2022). A Molecular Modeling Investigation of the Therapeutic Potential of Marine Compounds as DPP-4 Inhibitors. *Marine Drugs*, *20*(12), 777. https://doi.org/10.3390/md20120777
40. Zhao, Z., Ma, R., Ma, Y., Zhao, L., Wang, L., Fang, Y., Zhang, Y., Wu, X., & Wang, X. (2024). Discovery of Nine Dipeptidyl Peptidase-4 Inhibitors from Coptis chinensis Using Virtual Screening, Bioactivity Evaluation, and Binding Studies. *Molecules*, *29*(10), 2304. https://doi.org/10.3390/molecules29102304
41. Dăescu, D.-I., Păușescu, I., Benea, I. C., Peter, F., Todea, A., Zappaterra, F., Alexa, A. A., & Buzatu, A. R. (2024). Natural and Synthetic Flavylium Derivatives: Isolation/Synthesis, Characterization and Application. *Molecules*, *30*(1), 90. https://doi.org/10.3390/molecules30010090
42. Bayanati, M., Ismail Mahboubi Rabbani, M., Sirous Kabiri, S., Mir, B., Rezaee, E., & Tabatabai, S. A. (2024). Dipeptidyl Peptidase-4 Inhibitors: A Systematic Review of Structure-Activity Relationship Studies. *Iranian Journal of Pharmaceutical Research*, *23*(1). https://doi.org/10.5812/ijpr-151581
43. Mkabayi, L., Viljoen, Z., Krause, R. W. M., Lobb, K. A., Pletschke, B. I., & Frost, C. L. (2024). Inhibitory effects of selected cannabinoids against dipeptidyl peptidase IV, an enzyme linked to type 2 diabetes. *Heliyon*, *10*(1), e23289. https://doi.org/10.1016/j.heliyon.2023.e23289
44. Desantis, F., Miotto, M., Di Rienzo, L., Milanetti, E., & Ruocco, G. (2022). Spatial organization of hydrophobic and charged residues affects protein thermal stability and binding affinity. *Scientific Reports*, *12*(1), 12087. https://doi.org/10.1038/s41598-022-16338-5
45. Ayaz, M., Alam, A., Zainab, Elhenawy, A. A., Ahmad, I., Rahman, S. U., Ali, L., Latif, A., Ali, M., & Ahmad, M. (2026). Exploring the anti-diabetic potential of bis-Schiff bases of ibuprofen: insights into the *in vitro* , molecular docking and density functional theory analyses. *RSC Advances*, *16*(12), 11005–11022. https://doi.org/10.1039/D5RA03358F
46. Khalil, I., Zaman, S., Aliza, I. J., Rahman Tanny, M., Hasan, T., & Uzzaman, M. (2026). Structural tuning of diclofenac for enhanced medicinal efficacy and reduced adverse effects: an integrating computational validation. *RSC Advances*, *16*(26), 23740–23753. https://doi.org/10.1039/D6RA01053A
47. Ojeda-Montes, M. J., Gimeno, A., Tomas-Hernández, S., Cereto-Massagué, A., Beltrán-Debón, R., Valls, C., Mulero, M., Pujadas, G., & Garcia-Vallvé, S. (2017). Activity and selectivity cliffs for DPP-IV inhibitors: Lessons we can learn from SAR studies and their application to virtual screening. *Medicinal Research Reviews*, *38*(6), 1874–1915. https://doi.org/10.1002/med.21499
48. Baziyar, L., Solhjoo, A., Kamran, A., Mardaneh, P., Khabnadideh, S., Emami, M., Emami, L., & Rezaei, Z. (2025). Cytotoxic investigation of some Quinazoline-4 amino piperidine derivatives as DPP4 inhibitors along with molecular dynamic simulation studies. *Results in Chemistry*, *18*, 102779. https://doi.org/10.1016/j.rechem.2025.102779
49. Clemente-Suárez, V. J., Martín-Rodríguez, A., Beltrán-Velasco, A. I., Rubio-Zarapuz, A., Martínez-Guardado, I., Valcárcel-Martín, R., & Tornero-Aguilera, J. F. (2025). Functional and Therapeutic Roles of Plant-Derived Antioxidants in Type 2 Diabetes Mellitus: Mechanisms, Challenges, and Considerations for Special Populations. *Antioxidants*, *14*(6), 725. https://doi.org/10.3390/antiox14060725
50. Sarkar, D., Christopher, A., & Shetty, K. (2022). Phenolic Bioactives From Plant-Based Foods for Glycemic Control. *Frontiers in Endocrinology*, *12*. https://doi.org/10.3389/fendo.2021.727503



e-ISSN : 2622-4968, p-ISSN : 2622-1349 