Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Integrated XGBoost-Based Virtual Screening and Molecular Dynamics Simulation for the Identification of Potential CCR5 Antagonists

Angadi Sathish Kumar
Kakatiya University
MD Sanober
Kakatiya University
Estari Mamidala
drestari@kakatiya.ac.in

Kakatiya University

Research Article

Keywords: HIV, CCR5 receptor, Machine Learning, XGBoost, Molecular Docking, Molecular Dynamics Simulation

Posted Date: September 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10950316/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Integrated XGBoost-Based Virtual Screening and Molecular Dynamics Simulation for the Identification of Potential CCR5 Antagonists

First author:
Angadi Sathish Kumar[1]
Email: angadisathish1995@kakatiya.ac.in
ORCID ID: https://orcid.org/0009-0008-7771-8375

Second Author:
MD Sanober[1]
Email: sanobermd@kakatiya.ac.in
ORCID ID: https://orcid.org/0009-0005-8213-7805

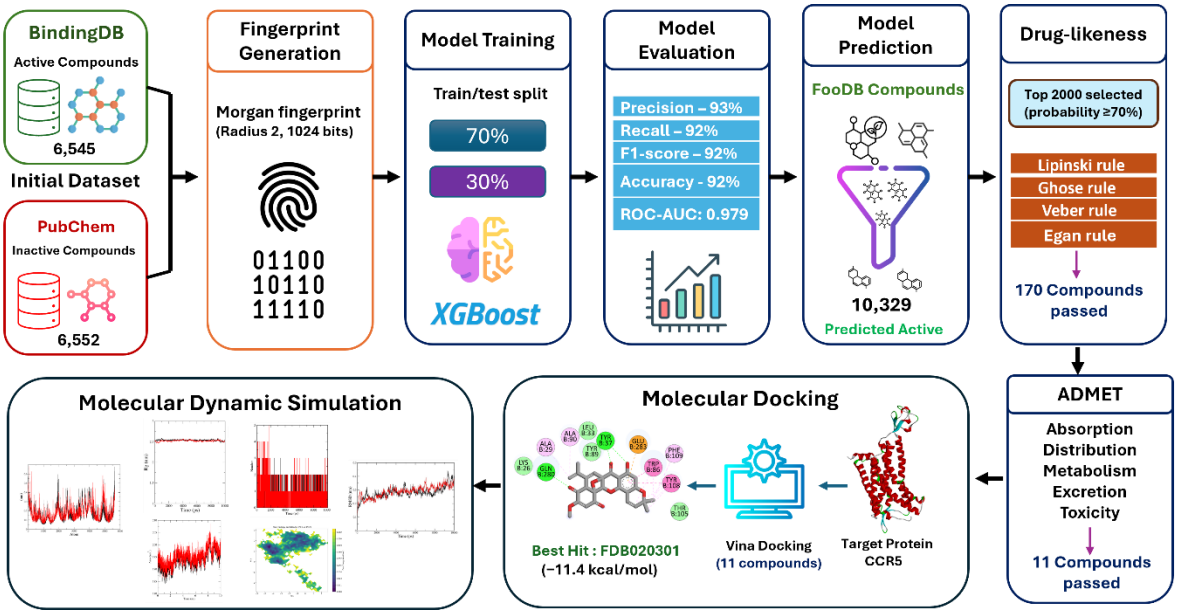
Corresponding author:
Prof. Estari Mamidala[1]
E-Mail: drestari@kakatiya.ac.in
ORCID ID: https://orcid.org/0000-0002-7874-6019

[1] Infectious Diseases Research Lab, Department of Zoology, Kakatiya University, Warangal-506009, Telangana State, India

Abstract

Human Immunodeficiency Virus (HIV) remains a major global health concern, highlighting the need for the continuous discovery of effective therapeutic agents targeting viral entry mechanisms. The C-C chemokine receptor type 5 (CCR5) is a key co-receptor involved in HIV entry into host cells and represents a promising target for antiviral drug development. Therefore, this study aimed to identify and prioritize potent CCR5 antagonists using an integrated machine learning-driven *in silico* drug discovery approach. A dataset of 170 drug-like molecules was subjected to predictive screening using an XGBoost (Extreme Gradient Boosting) ensemble learning model. The selected candidates were further evaluated for their pharmacokinetic and toxicity profiles through ADMET prediction. Subsequently, molecular docking simulations were performed using AutoDock Vina against the CCR5 receptor (PDB ID: 4MBS) to assess binding affinities and interaction patterns. The top-ranked docking hit was further subjected to molecular dynamics (MD) simulation. The resulting trajectories were subsequently analysed using principal component analysis (PCA). The docking analysis revealed strong binding affinities of the shortlisted compounds, ranging from −7.1 to −11.4 kcal/mol. Among them, FDB020301 exhibited the highest binding affinity (−11.4 kcal/mol), surpassing the clinically approved CCR5 antagonist Maraviroc (−10.5 kcal/mol). Molecular dynamics simulation further demonstrated the structural stability of the CCR5–FDB020301 complex, supporting the favourable binding mode and stability of the ligand within the receptor binding site. FDB020301 emerged as a potential lead molecule with superior binding affinity and favourable pharmacokinetic properties, representing a promising scaffold for next-generation anti-HIV drug development. However, further *in vitro* and *in vivo* studies are required to validate its antiviral efficacy and clinical applicability.

Keywords HIV CCR5 receptor Machine Learning XGBoost Molecular Docking Molecular Dynamics Simulation



Graphical Abstract

Introduction

The Human Immunodeficiency Virus (HIV) continues to be a recurring health issue worldwide even after the availability of antiretroviral therapy (ART) that has been very effective in prolonging the lives of patients [1]. The first phase of the HIV infection process involves a highly specialized interaction between the viral envelope protein gp120, the CD4 receptor, and the chemokine co-receptor, mostly the C-C chemokine receptor 5 (CCR5) [2,3]. The importance of CCR5 is highlighted by its ability to allow viral infection of immune cells such as macrophages and T lymphocytes, thus making it an ideal target in the treatment of HIV [4]. In terms of clinical utility, the significance of inhibiting the function of CCR5 has been proven via the use of Maraviroc, which is the first CCR5 antagonist approved for the treatment of HIV infections. Maraviroc works by interfering with the gp120 interaction with the cellular receptor CCR5 through the binding of Maraviroc to CCR5, causing conformational changes in the latter [5]. Although Maraviroc is successful in practice, a few concerns are associated with its use, such as development of drug resistance, low efficacy against CXCR4-tropic viruses [6].

Over the past few decades, there has been a significant development in drug discovery due to advances in the computational methodology, including Computer-Aided Drug Design (CADD). In this regard, machine learning (ML) algorithms have played an important role in modeling biological activity through the accurate and effective screening of massive chemical libraries [7]. Out of the many ML algorithms developed, Extreme Gradient Boosting (XGBoost) algorithm is gaining immense popularity due to its effectiveness, robustness, and better modeling capacity [8]. The combination of such ML methodologies with structural modeling approaches like molecular docking and pharmacokinetic ADMET prediction makes for highly efficient early-stage drug discovery processes [9,10]. Drug discovery is an elaborate process that can take several years and a lot of money to identify and develop a good lead compound. It is within this framework that *in silico* methods become an effective tool for discovering drug candidates that exhibit excellent bioactivity and pharmacokinetics. The use of both ligand-based and structure-based techniques can effectively accomplish this objective [11, 12].

The integration of ML-based virtual screening with natural-product-oriented chemical databases provides an attractive strategy for discovering new molecular scaffolds against therapeutically relevant targets such as CCR5. FooDB currently contains more than 70,000 listed compounds and incorporates information on chemical structures, biological activities, physicochemical characteristics, and food sources, creating an extensive chemical space for systematic computational screening [13]. Recent studies have demonstrated the feasibility of applying machine learning to FooDB-derived compounds for predicting health-associated biological activities, although the uneven distribution of activity annotations remains an important challenge for model development and validation [14]. Combining XGBoost-based prediction with subsequent pharmacokinetic and toxicity assessment can therefore facilitate the prioritization of compounds with favorable predicted biological and drug-like properties. Furthermore, structure-based methods such as molecular docking can provide complementary information by evaluating the potential binding modes and interactions of prioritized molecules within the CCR5 ligand-binding pocket. The availability of the experimentally determined CCR5–maraviroc structure provides a suitable structural framework for such investigations [15, 16]. Accordingly, the present study employed an integrated XGBoost-driven virtual screening strategy to interrogate the FooDB chemical space for potential high-affinity CCR5 inhibitors, followed by ADMET assessment and molecular docking to prioritize promising candidates for further computational and experimental investigation.

Materials and methods

Data Collection

The machine learning approach was employed to predict prospective CCR5 inhibitors using publicly accessible chemical and bioactivity databases. The initial dataset comprised 13,104 chemical compounds, including active CCR5 inhibitors (6545) obtained from BindingDB [17] and inactive compounds (6552) sourced from PubChem [18]. Consequently, a curated dataset of 13,097 compounds was obtained, comprising 6,545 active and 6,552 inactive compounds, and was used for subsequent machine learning model development.

Fingerprint generation

Simplified Molecular Input Line Entry System (SMILES) codes for all compounds (13,097) were processed using RDKit, which is a cheminformatics software library available in the public domain [19], to obtain chemical descriptors, and calculate Morgan fingerprints (with radius = 2, bits = 1024) [20] for encoding the structural information of compounds. All the machine learning algorithms used in this paper have been coded in Python (version 3.12.3) [21], in the Anaconda environment [22], under the Spyder IDE [23].

Model training

The generated molecular fingerprint dataset was randomly divided into a training set (70%) and an independent test set (30%). The training set was used to develop and optimize the XGBoost model, while the test set was reserved for evaluating its predictive performance on unseen compounds. A supervised XGBoost classifier algorithm was used for discrimination between active (1) and inactive (0) molecules because of its excellent predictive power and capability to detect non-linear interactions in chemical data [24]. The performance of the model was assessed based on different measures, such as accuracy, precision, recall, F1-score, Receiver Operating Characteristic (ROC) curve, and area under the curve (AUC) [25]. Feature selection was carried out in order to determine the significant fingerprint descriptors that influence the CCR5 inhibitor activity.

Model Prediction

The trained XGBoost classifier was subsequently applied to a large chemical library obtained from the FooDB database (https://foodb.ca/), comprising 69,606 compounds [26]. For each compound, circular molecular fingerprints (radius = 2; 1024 bits) were generated using the RDKit cheminformatics toolkit and used as input features for the trained XGBoost model. Based on the model predictions, 10,329 compounds were classified as potentially active against the target. To further prioritize the most promising candidates, the predicted active compounds were ranked according to their predicted probability of activity, and the 2,000 compounds with the highest predicted probabilities were selected for subsequent evaluation of drug-likeness properties. This sequential screening strategy substantially reduced the size of the initial FooDB chemical space while retaining compounds with a high predicted likelihood of biological activity. The prioritized compounds were then subjected to further computational analysis, including drug-likeness and ADMET profiling, before molecular docking studies against CCR5.

Virtual Screening

Drug-likeness properties

Following machine learning-based virtual screening, the 2,000 prioritized compounds were subjected to drug-likeness assessment to identify molecules with favorable physicochemical properties and potential suitability for oral administration. Drug-likeness was evaluated using the Lipinski, Ghose, Veber, and Egan filters through SwissADME (https://www.swissadme.ch/). Lipinski's Rule of Five evaluates key physicochemical parameters, including molecular weight, lipophilicity (LogP), hydrogen-bond donor count, and hydrogen-bond acceptor count, to identify compounds with properties generally associated with favorable oral bioavailability [27]. The Ghose, Veber, and Egan filters provide complementary criteria based on molecular size, lipophilicity, polar surface area, rotatable bonds, and related physicochemical characteristics [28, 29, 30]. Compounds satisfying the established criteria were retained for subsequent ADMET profiling and molecular docking studies against the CCR5 receptor.

ADMET Screening

The pharmacokinetic and safety profiles of the compounds that passed the drug-likeness filters were evaluated using Deep-PK (https://biosig.lab.uq.edu.au/deeppk/), a deep learning-based web platform for predicting the pharmacokinetic and toxicity properties of small molecules [31]. The assessment included key pharmacokinetic parameters related to absorption and distribution, including human intestinal absorption (HIA), blood–brain barrier (BBB) permeability, and interactions with drug transporters such as P-glycoprotein (P-gp) and organic cation transporter 2 (OCT2) [32]. Metabolic properties were evaluated based on the predicted interactions of the compounds with cytochrome P450 (CYP) enzymes, particularly their potential to act as substrates of CYP1A2 and inhibitors of CYP2D6. The toxicity profiles were assessed using predictions of Ames mutagenicity, carcinogenicity, and human ether-à-go-go-related gene (hERG) channel inhibition, the latter being an important indicator of potential cardiac toxicity [33]. Only compounds exhibiting favorable predicted pharmacokinetic characteristics, including good intestinal absorption and acceptable distribution and metabolic profiles, together with favorable safety predictions (non-mutagenic, non-carcinogenic, and low hERG inhibition potential), were prioritized for subsequent molecular docking studies against the CCR5 receptor.

Molecular Docking

Molecular docking studies have been conducted to examine the binding affinity of the 11 identified lead molecules towards the CCR5 receptor. The Crystal Structure of the CCR5 Chemokine Receptor (PDB ID 4MBS) was obtained from the Protein Data Bank (https://www.rcsb.org/structure/4MBS) [34]. The acquired protein structure was processed in BIOVIA Discovery Studio [35]. In the process of protein preparation, only chain A was retained, while all other chains, co-crystallized ligands, heteroatoms, and water molecules were eliminated to acquire a receptor structure for docking [36]. The addition of polar hydrogens and assigning the Kollman charges on the receptor molecule were performed to account for the electrostatic interactions. Finally, the prepared receptor structure was exported to the PDBQT file format to be used in the docking simulation studies.

For molecular docking studies, the chemical structure of the selected 11 lead compounds was used, which have been determined through virtual screening using the machine learning approach (**Fig. 1**). These chemical structures were taken in the form of SMILES. The obtained SMILES were used to generate three-dimensional structure and saved in the form of SDF file with the help of open babel [37]. Ligands' charge was set to Gasteiger charge and rotatable bonds were specified for docking studies. Then, optimized structures were converted to the format required for docking studies that is PDBQT file format.

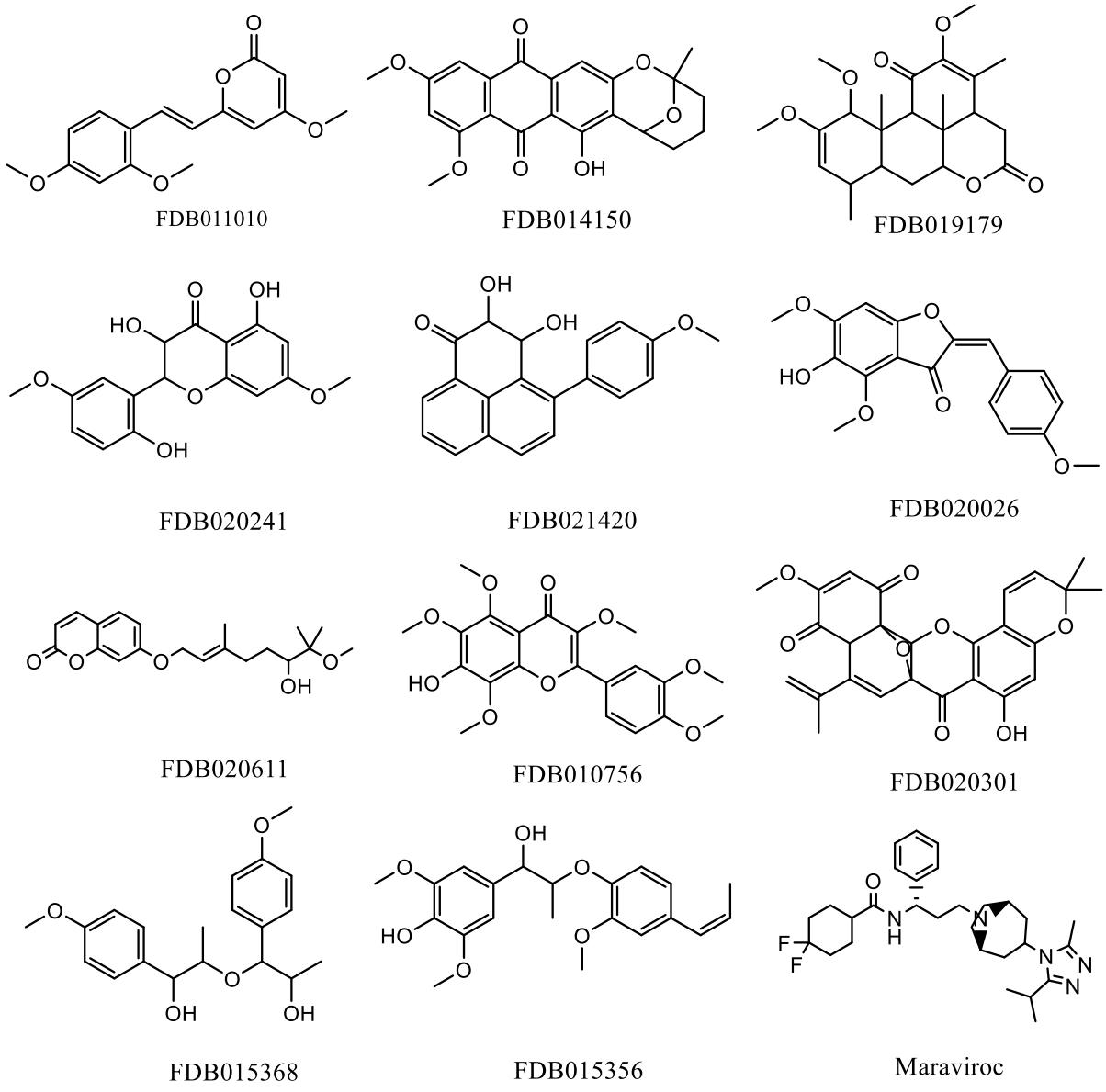


Fig. 1 2D structures of eleven filtered drug compounds and control drug (Maraviroc)

The molecular docking simulation was carried out utilizing AutoDock Vina for prediction of binding modes and affinities of 11 lead compounds in the binding site of the CCR5 receptor. The pre-modeled receptor and grid box (center coordinates at x=188.654, y=112.908, z=22.935 Å; dimensions: 90 × 70 × 70 Å) were applied for docking. Parameters used for docking include exhaustiveness = 32, number of binding modes = 9, and energy range = 3 kcal/mol. Binding affinity (ΔG, kcal/mol) was applied to rank the binding modes and $\Delta G \leq -8.0$ kcal/mol for

further analysis. For interaction studies, the 2D scheme diagrams were generated using BIOVIA Discovery Studio Visualizer software package.

Dynamic Simulation

RMSD, RMSF and Rg analysis

Molecular dynamics (MD) simulations were performed using GROMACS 2023.2 to investigate the structural stability and dynamic behavior of the CCR5–FDB020301 complex [38]. The protein–ligand complex was parameterized using the CHARMM36 all-atom force field, while ligand topology parameters were generated using the CHARMM General Force Field (CGenFF). The complex was solvated in a suitable water box, and Na^+ and Cl^- ions were added to neutralize the system. Energy minimization was performed using the steepest-descent algorithm to eliminate unfavourable contacts. The minimized system was equilibrated under NVT conditions for 100 ps at 300 K using the V-rescale thermostat, followed by NPT equilibration for 100 ps at 1 bar using the Parrinello–Rahman barostat. The equilibrated system was subsequently subjected to the production molecular dynamics simulation.

The stability and conformational behavior of the CCR5–FDB020301 complex were evaluated by calculating the root mean square deviation (RMSD) of the protein backbone and ligand–protein complex, root mean square fluctuation (RMSF) of individual residues, and radius of gyration (Rg) of the protein throughout the simulation trajectory [39]. RMSD analysis was used to assess overall structural stability, RMSF analysis to characterize residue-level flexibility, and Rg analysis to monitor changes in the compactness of the protein structure. Trajectory processing, analysis, and visualization were performed using GROMACS analysis utilities and Xmgrace.

Hydrogen-bond analysis

Hydrogen-bond analysis was performed to characterize the stability of polar interactions between CCR5 and FDB020301 throughout the 10 ns molecular dynamics simulation. MD trajectory using the GROMACS gmx hbond utility. They were identified according to geometric criteria based on the donor–acceptor distance and hydrogen–donor–acceptor angle. A donor–acceptor distance cutoff of 0.35 nm and an angular cutoff of 30° were used to define hydrogen-bond formation, consistent with the criteria implemented in GROMACS. The number of protein–ligand hydrogen bonds were monitored as a function of simulation time, and the mean number and range of hydrogen bonds were determined over the complete 10 ns trajectory [40].

SASA Analysis

Solvent-accessible surface area (SASA) analysis was performed to evaluate changes in solvent exposure and structural compactness of CCR5 upon binding of FDB020301 during the 10 ns molecular dynamics simulation. using GROMACS. The calculation was performed separately for the apo CCR5 protein, FDB020301 ligand, and the CCR5–FDB020301 complex using a solvent probe radius of 0.14 nm [41]. The buried solvent-accessible surface area at the protein ligand interface was estimated from the difference between the individual and complex SASA values according to:

$$SASA_{\text{buried}} = SASA_{\text{protein}} + SASA_{\text{ligand}} - SASA_{\text{complex}}$$

Lower and relatively stable SASA values were interpreted as indicative of reduced solvent exposure and maintenance of structural compactness.

PCA analysis

Principal component analysis (PCA) was performed using the GROMACS utilities to characterize the dominant collective motions and conformational fluctuations of the CCR5–FDB020301 complex during the 10 ns MD simulation. The covariance matrix was diagonalized to obtain eigenvalues and eigenvectors, and the contribution of individual principal components to the total variance was determined. The trajectory was projected onto PC1 and PC2 to evaluate the conformational space sampled by the complex. A two-dimensional free-energy landscape (FEL) was subsequently constructed from the PC1–PC2 distribution using the Boltzmann relationship, allowing identification of the preferred low-energy conformational states and assessment of conformational stability [42].

MM-GBSA calculation

Binding free energy calculations were performed using the Molecular Mechanics–Generalized Born Surface Area (MM-GBSA) methods implemented in the module of AMBER. Solvent molecules and counterions were removed prior to energy calculations [40]. For each trajectory, 500 representative frames were extracted from 20,000 total trajectory frames at a sampling interval of 40 frames from equilibrated simulation regions. MM-GBSA calculations were performed using the generalized Born implicit solvent model with igb = 2 under a salt concentration of 0.15 M. All binding energies are reported in kcal/mol as mean ± standard deviation [43].

Results

Machine Learning-Based Prediction of CCR5 Inhibitors

Model Performance Evaluation

The performance of the machine learning algorithm was measured using several evaluation metrics, including accuracy, precision, recall, and F1-score. Additionally, the classification performance was analyzed using confusion matrices and the ROC curve . The machine learning model developed using the XGBoost algorithm was evaluated on an independent test dataset comprising 13,097 chemical compounds, including 6,552 inactive compounds (Class 0) and 6,545 active compounds (Class 1). The model achieved an overall accuracy of 92%, indicating its strong predictive performance. The classification report further demonstrated good predictive capability. For Class 0 (inactive compounds), the model achieved a precision of 0.88, recall of 0.97, and F1-score of 0.93, indicating that most inactive compounds were correctly identified. Similarly, for Class 1 (active compounds), the model achieved a precision of 0.97, recall of 0.87, and F1-score of 0.92, demonstrating the model's effectiveness in predicting potential CCR5 inhibitor compounds (**Table 1**).

Table 1 Model performance evaluation

Metric	Active (1)	Inactive (0)	Macro Avg	Weighted Avg
Precision	0.97	0.88	0.93	0.93
Recall	0.87	0.97	0.92	0.92
F1-score	0.92	0.93	0.92	0.92
Support	6545	6552	–	–
Accuracy	-	-	-	92

ROC Curve Analysis

To determine the discriminating power of the proposed model, the receiver operating characteristic (ROC) curve for the same was plotted using Python (Version 3.12) with the help of the scikit-learn and matplotlib packages. The ROC of the model yielded a value of 0.979, reflecting high discrimination ability of the model between the active and the inactive chemical compounds. The value of AUC being near to one shows that the discriminating power of the proposed model is very high (**Fig. 2**).

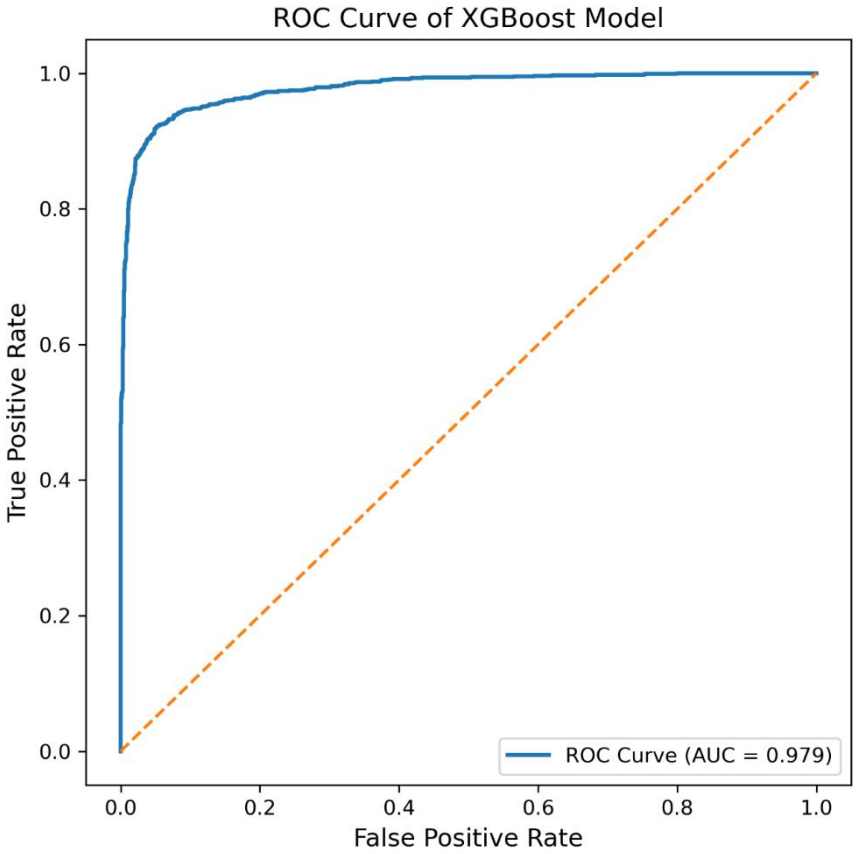


Fig. 2 Receiver Operating Characteristic (ROC) curve of the XGBoost classification model showing an AUC value of 0.979 for distinguishing active and inactive CCR5 inhibitors

Confusion Matrix Analysis

The confusion matrix provides additional insights into the performance of the classifier model by displaying the numbers of true positives, false positives, true negatives, and false negatives. The model was able to classify 6,383 inactive compounds (true negatives) and 5,710 active compounds (true positives) accurately. There were only few instances where the compounds were not classified correctly. Inactive compounds that were classified as active (false positives) totalled 169, while active compounds classified as inactive (false negatives) reached 835 (**Fig. 3**).

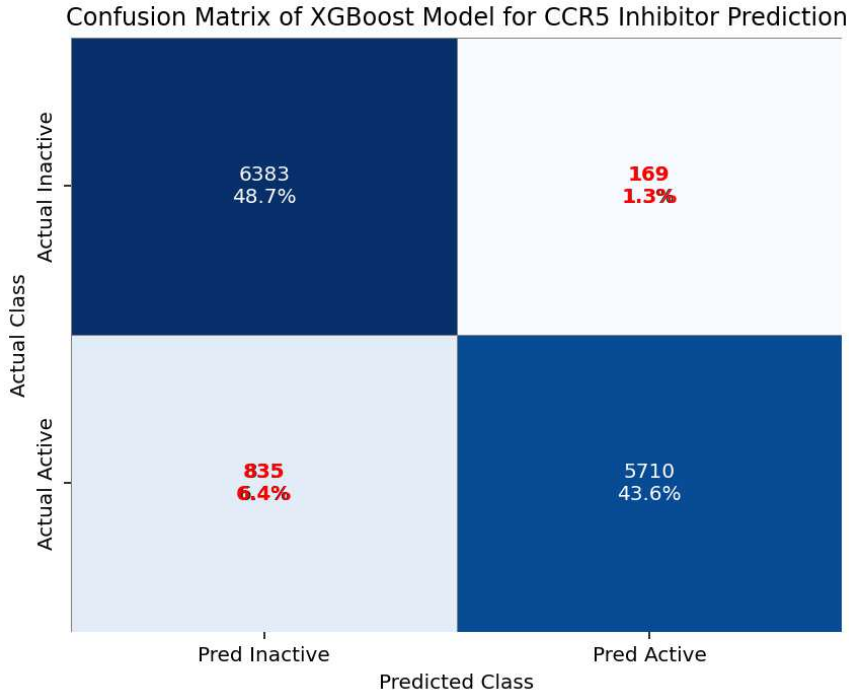


Fig. 3 Confusion matrix illustrating the classification performance of the XGBoost model for prediction of CCR5 inhibitors

Feature Importance Analysis

To determine which molecular fingerprint features contributed most to the prediction of the CCR5 inhibition activity, a feature importance analysis was conducted. The feature importance values for the descriptors obtained from the Morgan fingerprints using RDKit in Python were plotted to find out which descriptors played important roles in predicting the CCR5 inhibition activity. The **Fig. 4** shows the top fingerprint features in terms of the contributions to the classification model, suggesting that certain substructure patterns in molecules play important roles in CCR5 inhibition activity prediction.

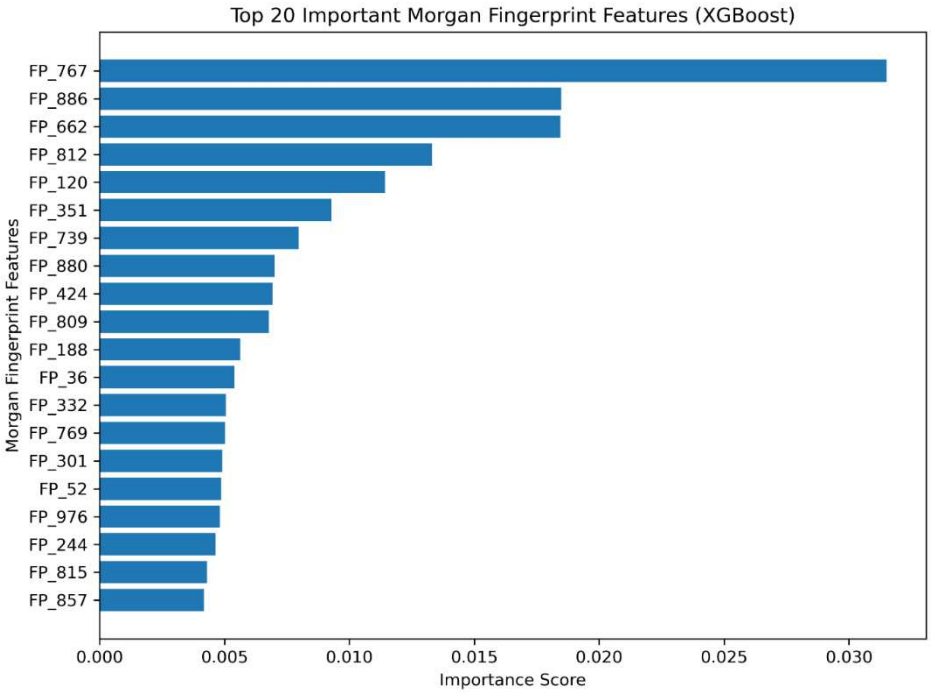


Fig. 4 Feature importance plot showing the most influential Morgan fingerprint descriptors contributing to CCR5 inhibitor prediction

External Validation

As an additional step to measure the prediction capabilities of the trained model, external validation was conducted with the help of CCR5 targeting drugs. The trained model was validated against FDA approved and investigational molecules that were observed to bind with the target receptor. This includes Maraviroc, a clinically approved CCR5 inhibitor. The prediction confidence score for these molecules was found to be high (~0.99). This indicates the effectiveness and accuracy of the trained model in predicting CCR5 binding molecules.

Model Prediction Outcome

The trained XGBoost model was applied to the pre-processed FooDB dataset to classify compounds as predicted active or inactive based on their probability scores. A probability threshold of 0.70 (70%) was used to identify compounds with a high likelihood of CCR5 inhibitory activity. Of the 69,606 compounds remaining after data cleaning and preprocessing, 10,329 compounds were predicted to be potentially active against CCR5. These compounds were subsequently ranked according to their predicted activity probabilities, and the 2,000 compounds with the highest probability scores (≥70%) were prioritized for further drug-likeness evaluation.

Virtual Screening of potential drugs

Drug-likeness analysis

The drug-likeness analysis was performed using the Lipinski, Ghose, Veber, and Egan filters to evaluate the physicochemical suitability of the screened compounds. The selected compounds were assessed based on molecular weight, lipophilicity, hydrogen-bond donors and

acceptors, molar refractivity, total atom count, rotatable bonds, and topological polar surface area. As summarized in **Table 2**, the retained compounds satisfied the criteria of all four drug-likeness rules, indicating favorable physicochemical properties and potential suitability for oral administration. Overall, of the 2000 compounds screened, 170 compounds successfully passed all the applied filters and were subsequently prioritized for ADMET profiling and further computational evaluation against the CCR5 receptor (Supplementary table S1).

Table 2 Comprehensive drug-likeness evaluation of screened compounds using multiple physicochemical filtering criteria

Sl.No	Drug-likeness Rule	Criteria Parameters	Criteria Satisfied YES / NO
1	Lipinski Rule	Molecular weight ≤ 500 Da; LogP ≤ 5; H-bond donors ≤ 5; H-bond acceptors ≤ 10	YES
2	Ghose Filter	Molecular weight 160–480; LogP −0.4 to 5.6; Molar refractivity 40–130; Total atoms 20–70	YES
3	Veber Rule	Rotatable bonds ≤ 10; Topological polar surface area (TPSA) ≤ 140 Å²	YES
4	Egan Rule	LogP ≤ 5.88; TPSA ≤ 131 Å²	YES
5	-	Compounds satisfying all drug-likeness criteria	170

ADMET profile analysis

Among the 170 compounds that passed the drug-likeness filters, 11 compounds demonstrated favorable predicted ADMET profiles and were selected for further evaluation. As presented in **Table 3**, all 11 selected compounds were predicted to exhibit human intestinal absorption, indicating favorable absorption potential. All compounds were predicted to act as P-glycoprotein inhibitors and CYP1A2 substrates, while none were predicted to inhibit CYP2D6. Regarding distribution, six compounds were predicted to be blood–brain barrier penetrable, whereas five compounds, including FDB020301, were predicted to be non-penetrable. Overall, these results indicate favorable predicted pharmacokinetic characteristics for the shortlisted compounds.

Table 3 Predicted Pharmacokinetic Properties of Selected Lead Compounds

S. N	Compound ID	Human Intestinal Absorption	P-Glycoprotein Inhibitor/Non-inhibitor	Blood-Brain Barrier	CYP 1A2 Substrate \non-substrate	CYP 2D6 Inhibitor/Non-inhibitor
1	FDB011010	Absorbed	Inhibitor	Penetrable	Substrate	Non-Inhibitor
2	FDB014150	Absorbed	Inhibitor	Penetrable	Substrate	Non-Inhibitor
3	FDB019179	Absorbed	Inhibitor	Penetrable	Substrate	Non-Inhibitor
4	FDB020241	Absorbed	Inhibitor	Non-Penetrable	Substrate	Non-Inhibitor

5	FDB021420	Absorbed	Inhibitor	Non-Penetrable	Substrate	Non-Inhibitor
6	FDB020026	Absorbed	Inhibitor	Penetrable	Substrate	Non-Inhibitor
7	FDB020611	Absorbed	Inhibitor	Penetrable	Substrate	Non-Inhibitor
8	FDB010756	Absorbed	Inhibitor	Non-Penetrable	Substrate	Non-Inhibitor
9	FDB020301	Absorbed	Inhibitor	Non-Penetrable	Substrate	Non-Inhibitor
10	FDB015368	Absorbed	Inhibitor	Non-Penetrable	Substrate	Non-Inhibitor
11	FDB015356	Absorbed	Inhibitor	Non-Penetrable	Substrate	Non-Inhibitor

As depicted in **Table 4**, all 11 selected compounds exhibited favorable predicted transporter and toxicity profiles. All compounds were predicted to be non-inhibitors of organic cation transporter 2 (OCT2) and showed safe predictions in the AMES mutagenicity and carcinogenicity assessments. Furthermore, none of the selected compounds were predicted to exhibit hERG channel-blocking activity, suggesting a low potential for hERG-associated cardiotoxicity. Collectively, the favorable predicted safety profiles supported the selection of these 11 compounds, including FDB020301, for subsequent molecular docking and further computational analyses.

Table 4 Predicted Toxicological and Transporter Profiles of Selected Lead Compounds

S. N	Compound ID	Organic Cation Transporter 2	AMES Mutagenesis	Carcinogenesis	hERG Blockers
1	FDB011010	Non-Inhibitor	Safe	Safe	Safe
2	FDB014150	Non-Inhibitor	Safe	Safe	Safe
3	FDB019179	Non-Inhibitor	Safe	Safe	Safe
4	FDB020241	Non-Inhibitor	Safe	Safe	Safe
5	FDB021420	Non-Inhibitor	Safe	Safe	Safe
6	FDB020026	Non-Inhibitor	Safe	Safe	Safe
7	FDB020611	Non-Inhibitor	Safe	Safe	Safe
8	FDB010756	Non-Inhibitor	Safe	Safe	Safe
9	FDB020301	Non-Inhibitor	Safe	Safe	Safe
10	FDB015368	Non-Inhibitor	Safe	Safe	Safe
11	FDB015356	Non-Inhibitor	Safe	Safe	Safe

Molecular Docking Analysis

Molecular docking studies were carried out to assess the binding affinity and binding pattern of the 11 lead molecules in the binding site of the CCR5 receptor (PDB ID: 4MBS). All the shortlisted molecules showed good binding with the receptor, having docking energy values between −7.1 and −11.4 kcal/mol, suggesting successful formation of the protein-ligand complex. These findings were then compared with the FDA-approved CCR5 inhibitor Maraviroc, whose binding energy value was −10.5 kcal/mol (**Table 5**).

Table 5 Molecular Docking Scores and Interaction Profiles with CCR5 Receptor

S. N	Compound ID FooDB	Binding Energy (kcal/mol)	Number of Hydrogen bonds	Number of Hydrogen Amino acids	Other interactive amino acids
1	FDB011010	-7.1	1	TYR37	TYR89, ALA90, TRP86, TYR108, THR284, MET287, GLU283
2	FDB014150	-9.4	1	TYR37	GLU283, TYR108, TRP86, MET287, LEU33, ALA29, LYS26, TYR89, GLN280, ALA90, CYS178, THR105, PHE109
3	FDB019179	-7.7	1	THR167	TYR37, GLU283, TYR89, CYS178, TRP86, SER179, TYR108, THR105, PHE109, SER180, ASN163
4	FDB020241	-8.2	2	CYS178, THR167	GLU283, TRP86, TYR251, PHE109, THR105, SER179, TYR108, TYR89, TYR37, SER180
5	FDB021420	-9.3	2	TYR108, TYR37	CYS178, TRP86, TYR89, GLN280, GLU283, MET287, ALA90, LEU33, ALA29, LYS26
6	FDB020026	-7.7	1	SER180	THR105, GLN194, ASN163, PHE182, PHE109, TYR108, TRP86, GLU283, TYR251, MET279, MET287, TYR37

7	FDB020611	-8.8	1	TYR251	TYR37, MET287, TRP86, GLU283, TYR108, THR105, PHE109, ILE198, PHE113, PHE112, TRP248
8	FDB010756	-7.1	1	TYR37	GLU283, TYR108, TRP86, CYS178, TYR89, GLN280, ALA90, ALA29, THR284, LEU33, MET287
9	FDB020301	-11.4	2	TYR37, GLN280	LYS26, ALA29, ALA90, LEU33, TYR89, GLU283, TRP86, PHE109, TYR108, THR105
10	FDB015368	-7.4	2	THR105, GLU283	PHE109, PHE112, ILE198, TYR251, SER180, SER179, THR167, TYR37, TRP86, TYR108
11	FDB015356	-8.6	-	-	TRP248, TYR251, PHE112, ILE198, PHE109, GLU283, SER283, SER180, MET287, TRP86, TYR108, TYR37, TYR89, THR105, MET279
12	Maraviroc (FDA approved drug)	-10.5	1	TYR251	PHE112, ILE198, TRP248, TYR108, TRP86, TYR89, GLU283, TYR37, LEU33, ALA90, GLN280, THR284, MET279, PHE182, SER180, PHE109

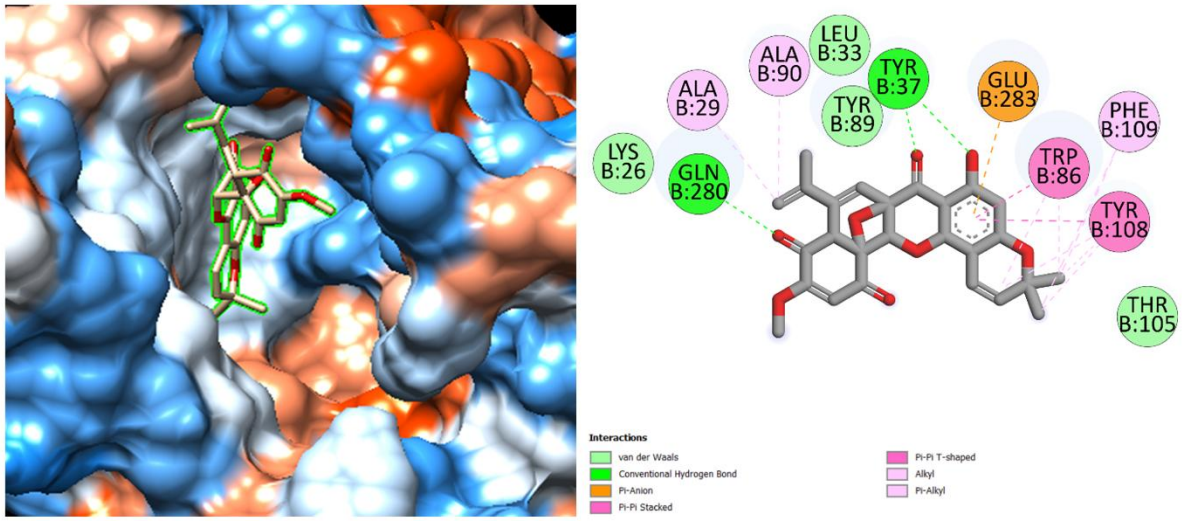


Fig. 5 3D surface representation and 2D visualization of molecular docking results of FDB020301 with Crystal Structure of the CCR5 Chemokine Receptor (PDB ID 4MBS)

Among the tested compounds, FDB020301 showed the highest binding affinity (−11.4 kcal/mol), exceeding that of the reference drug. It formed hydrogen bonds with TYR37 and GLN280, along with Pi-Anion and Pi-Pi interactions involving GLU283, TRP86, and TYR108. Additional hydrophobic interactions with TYR89, PHE109, ALA29, ALA90, and PHE109 contributed to its binding within the CCR5 pocket (**Fig. 5**). The overlap of key interacting residues with those involved in Maraviroc binding supports FDB020301 as a promising candidate for further CCR5 inhibitor investigation.

Molecular Dynamic Simulation

RMSD analysis

A 10 ns molecular dynamics (MD) simulation was performed using GROMACS 2023.2 to evaluate the dynamic stability of FDB020301 within the CCR5 binding pocket. The protein RMSD initially increased and subsequently stabilized within a defined range of 0.0005–0.3927 nm, with a mean value of 0.2884 ± 0.0414 nm. During the final 5 ns, the mean RMSD was 0.3176 nm, with the maximum deviation (0.3927 nm) observed at approximately 7.09 ns. The ligand RMSD ranged from 0.0005–0.4128 nm, with a mean of 0.2105 ± 0.0470 nm. A transient peak of 0.4128 nm occurred at 4.65 ns, followed by stabilization, with a mean RMSD of 0.2030 ± 0.0427 nm during the final 5 ns (**Fig. 6**). The protein and ligand RMSD profiles indicated a stable CCR5–FDB020301 complex throughout the simulation, with the transient ligand deviation likely reflecting local readjustment within the binding pocket.

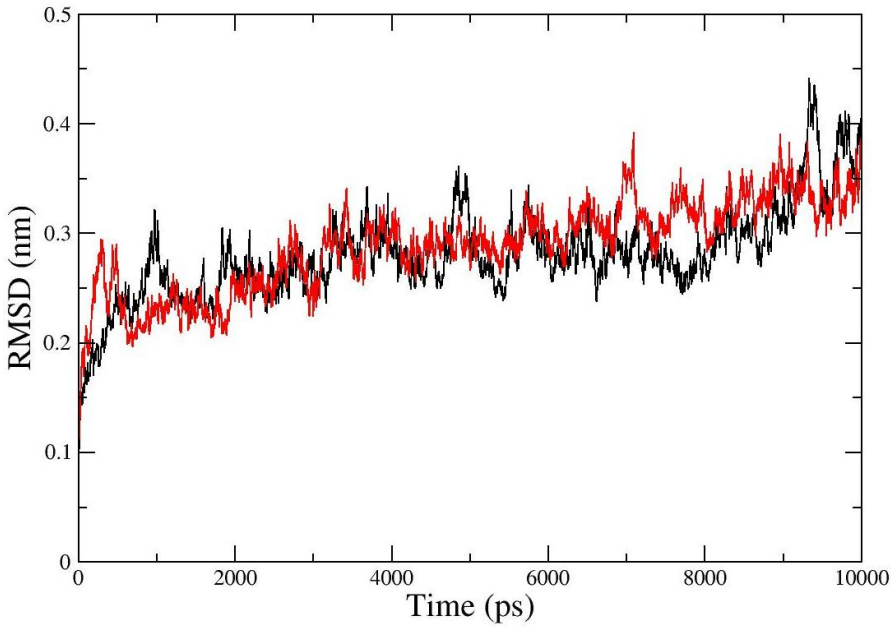


Fig. 6 RMSD analysis of the CCR5 protein (block colour) and FDB020301 ligand (Red colour) demonstrating the dynamic stability of the CCR5–FDB020301 complex during the 10 ns molecular dynamics simulation

RMSF Analysis

Based on Cα RMSF, the mean RMSF value was 0.1446 ± 0.0673 nm, with an RMSF range of 0.0608-0.4065 nm (0.61-4.07 Å) (**Fig. 7**). There was a tendency for most residues to be relatively stable, but residues with high fluctuations occurred only in particular sites within CCR5, suggesting local flexibility without affecting the overall stability of the structure. It is apparent that the predefined residues in the ligand binding site are relatively stable in the 10 ns molecular dynamics simulation.

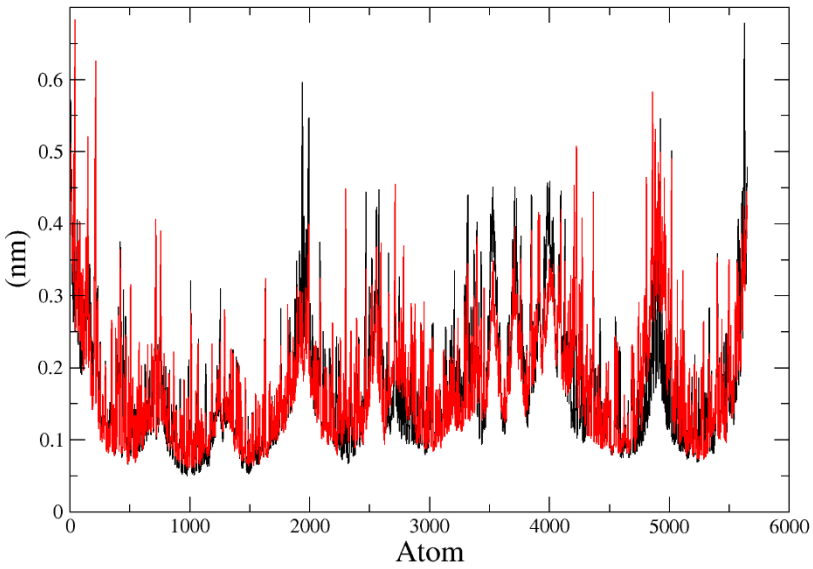


Fig. 7 RMSF analysis of the CCR5 protein (Black colour) and FDB020301 ligand (Red colour) demonstrating the dynamic flexibility of the CCR5–FDB020301 complex during the 10 ns molecular dynamics simulation

Radius of gyration

The Radius of Gyration (Rg) analysis was conducted to evaluate the compactness and structural stability of the CCR5 receptor and its complex with FDB020301 over the 10 ns molecular dynamics simulation. The Rg value represents the mass-weighted root mean square distance of atoms from their center of mass and serves as an indicator of the protein's tertiary structure integrity during the simulation. The radius of gyration was almost constant during the whole 10 ns MD simulation with an average value of 2.5232 ± 0.0171 nm and a minimum-maximum range of 2.4561-2.5770 nm. Over the last 5 ns, the average radius of gyration was 2.5225 ± 0.0142 nm with minimum and maximum values of 2.4779 and 2.5678 nm, respectively (**Fig. 8**). It can be noted that there is a strong similarity between overall and last 5 ns average values of radius of gyration, indicating no significant alteration in the compactness of the receptor.

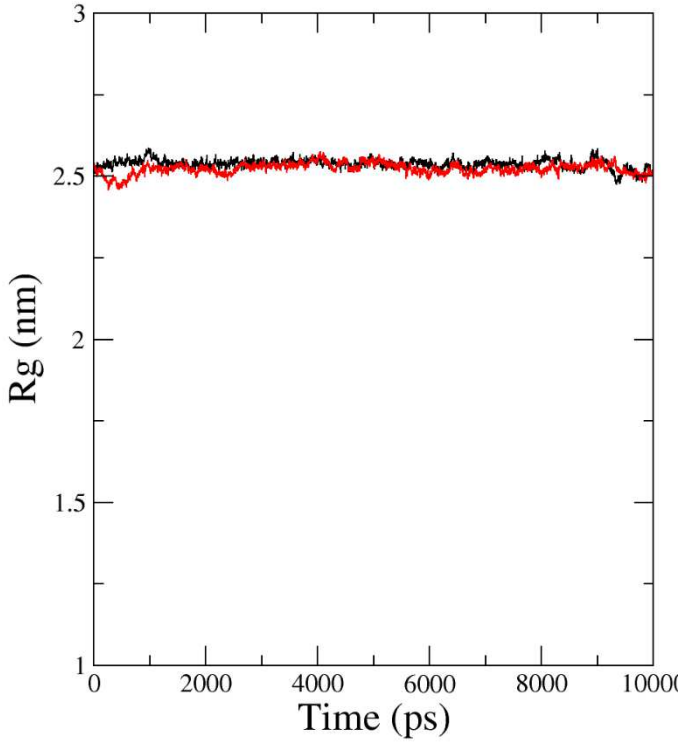


Fig. 8 Rg analysis of the CCR5 protein (Black colour) and FDB020301 ligand (Red colour) demonstrating the dynamic compactness of the CCR5–FDB020301 complex during the 10 ns molecular dynamics simulation

Hydrogen bond analysis

Analysis of hydrogen bonding was carried out to study the stability of polar interactions between CCR5 and FDB020301 during the molecular dynamics simulation. On average, there were 0.93 ± 0.48 hydrogen bonds between the complexes, with a range of 0 to 3 hydrogen bonds during the entire 10 ns time span. In the last 5 ns time, there were 0.86 ± 0.50 hydrogen bonds on average, with a range of 0 to 2 hydrogen bonds (**Fig. 9**). Thus, the stability of hydrogen bonds during the simulations suggests that polar interactions stabilized the complex.

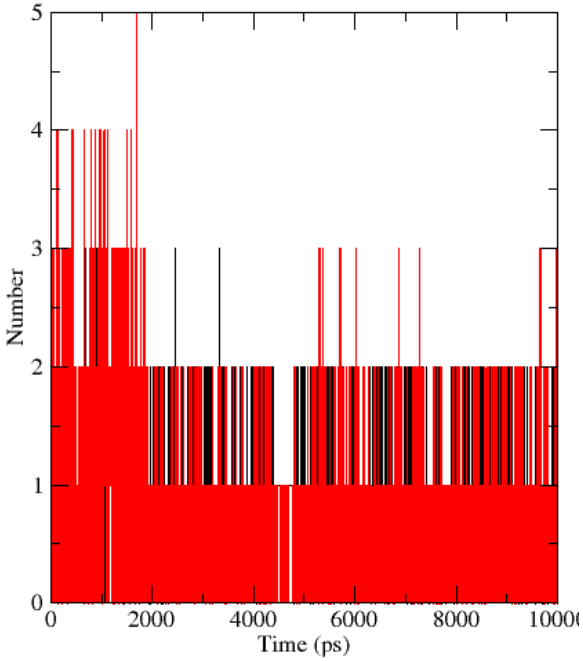


Fig. 9 Hydrogen bonding occupancy of the protein ligand complex over the 10ns molecular dynamic simulation

SASA Analysis

Solvent-accessible surface area (SASA) is another key property which can be employed in the determination of solvent accessibility of a protein and analysis of changes in its structural compactness during molecular dynamics simulations. Typically, the lower SASA values correspond to smaller solvent exposure and indicate a more compact and structurally stable protein structure. In the current study, the CCR5 protein had a mean SASA of 195.050 nm² and a range of 185.877–208.597 nm² within the 10 ns MD run. For FDB020301, the mean SASA was 6.218 nm², while the complex formed by CCR5 and FDB020301 had a mean SASA of 193.989 nm² (184.027–207.564 nm²). The somewhat lower mean SASA value of the complex compared to that of the protein suggests a minor reduction in solvent exposure after ligand binding. Moreover, the estimated mean buried SASA of the interface of protein and ligand amounted to 7.279 nm² (72.79 Å²), implying efficient burial of the interface area (**Fig. 10**). Taken together, the moderate difference in SASA between the protein and complex indicates lack of significant expansion or disruption of CCR5 upon ligand binding.

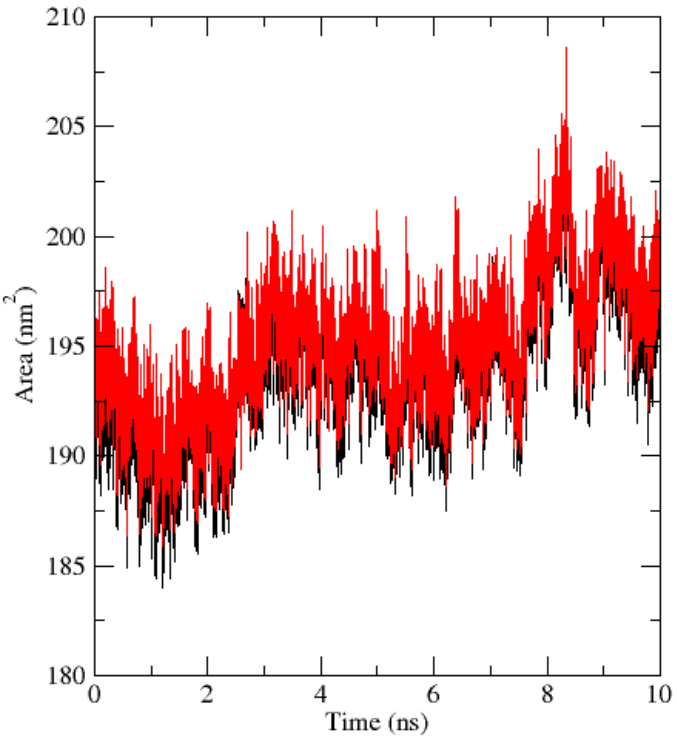


Fig. 10 SASA plots of CCR5 protein (in black), CCR5 complex (in red)

Principal component analysis

The principal component analysis (PCA) was used to analyze the main collective modes of motion of the protein-ligand complex during the 10 ns molecular dynamics simulation. According to the results of the covariance matrix analysis, the total variance was equal to 8.7984 nm², the variance of the first principal component (PC1) was 2.8796 nm² (32.73%), while the variance of the second principal component (PC2) was 1.7664 nm² (20.08%). Thus, the first two principal components explained 52.81% of the total variance, while the first three PC explained 64.94%, suggesting that a significant part of the overall conformational variability was captured by the principal modes at the very beginning of the calculation. Two-dimensional PCA plot of PC1 against PC2 shows the sampling of the corresponding conformational space by the trajectory of the complex, which occupied a few different but related regions of the structure rather than exhibited an unrestricted variability of the structures. To obtain additional information on the energy distribution of the conformational states obtained by PCA, the free-energy landscape (FEL) for PC1 and PC2 was constructed. As seen from the plot, the FEL has several regions of low free energy, the lowest-energy conformational basin is located approximately around PC1 ≈ −4.5 and PC2 ≈ 1.4, with the other notable low energy region lying between PC1 ≈ −4.0 to −3.0 and PC2 ≈ 0–0.5. The studied conformational space extends roughly from −5.8 to 2.2 in PC1 and from −2.7 to 3.1 in PC2, with the relative free-energy range being 0–1.9 kcal/mol (**Fig. 11**). Therefore, the analysis of PCA combined with FEL suggests the presence of a few low-energy conformations sampled by the complex and thus its overall dynamical stability.

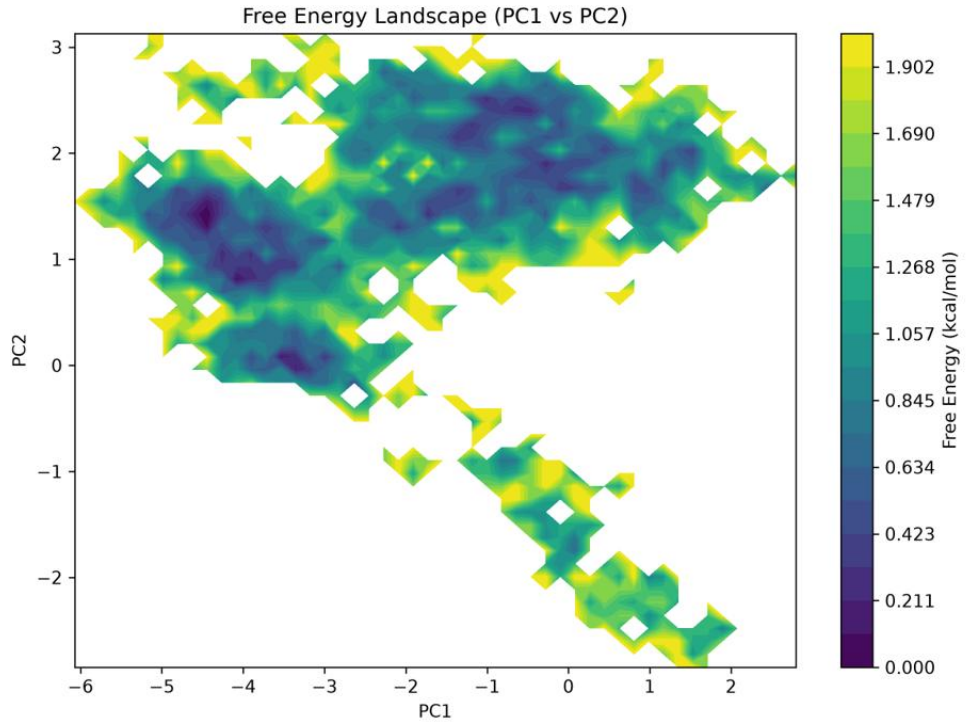


Fig. 11 Principal component analysis (PCA) of CCR5 complex during 10ns molecular dynamics simulations

MM/GBSA Binding free energy calculation

The binding energetics of the HIV-1 CCR5–complex was examined further using the MM/GBSA method. This analysis used 501 frames taken from 10-nanosecond molecular dynamics run. The calculated binding free energy, ΔG_bind was −35.19 ± 2.94 kcal mol⁻¹ showing that binding is energetically favorable. The gas-phase contribution, ΔG_GAS was −50.80 ± 5.51 kcal mol⁻¹ while the solvation contribution, ΔG_SOLV was +15.60 ± 3.86 kcal mol⁻¹. Looking at the parts van der Waals interactions gave the strongest favorable effect ΔVDW = −45.43 ± 2.81 kcal mol⁻¹. Electrostatic interactions were also favorable ΔEEL = −5.36 ± 4.77 kcal mol⁻¹ and nonpolar solvation helped, ΔESURF = −4.56 ± 0.39 kcal mol⁻¹. However, the polar-solvation term, ΔEGB = +20.16 ± 3.94 kcal mol⁻¹, worked against these interactions. Per-residue analysis showed that TRP86 contributed −5.76 ± 0.24 kcal mol⁻¹ TYR89 contributed −2.79 ± 0.29 kcal mol⁻¹ TYR108 contributed −1.98 ± 0.29 kcal mol⁻¹ TYR37 contributed −1.69 ± 0.22 kcal mol⁻¹ and ALA90 contributed −1.23 ± 0.17 kcal mol⁻¹. These residues are the favorable partners for the ligand. The energetic role of TRP86 matches the PLIP analysis, which found that TRP86 made contacts in 99.40 % of the simulation frames and π-stacking contacts in 22.95 % of the frames (**Table 6**). Likewise, TYR108, TYR89, TYR37 and THR105 kept PLIP contacts and gave good per-residue energies confirming their importance, in holding the ligand in place.

Table 6 MM-GBSA binding free energy profile of the FDB020301-CCR5 complex

Energy component	MEAN	SD
ΔEvdW	-45.4361	2.812447
ΔEele	-5.36926	4.772644
ΔEGB	20.16966	3.948609
ΔEsurf	-4.56265	0.394011
ΔGGas	-50.8055	5.517571
ΔGSolv	15.60697	3.869445
ΔGbind	-35.1983	2.940762

Discussion

The present study demonstrates the utility of machine learning-assisted screening for efficiently prioritizing potential CCR5 inhibitors from a large food-derived chemical library. The XGBoost model was applied to 69,606 pre-processed FooDB compounds, resulting in the identification of 10,329 compounds predicted to be active, from which the 2,000 highest-probability candidates were prioritized for subsequent screening. This progressive reduction of chemical space illustrates the advantage of ML-based approaches in rapidly focusing computational resources on compounds with a greater predicted likelihood of biological activity. XGBoost is particularly suitable for classification problems because it employs gradient-boosted decision trees to capture complex and nonlinear relationships between molecular features and biological outcomes, while maintaining flexibility and computational efficiency. Recent studies have also emphasized the growing role of XGBoost and related AI/ML approaches in drug discovery and development, while highlighting the importance of appropriate validation and cautious interpretation of model predictions [44]. Thus, the present predictions should be considered as a prioritization strategy rather than direct evidence of CCR5 inhibition. Nevertheless, the integration of molecular fingerprints with XGBoost enabled systematic interrogation of the large FooDB chemical space and generated a manageable set of candidates for downstream drug-likeness, ADMET, and structure-based analyses. The screening outcome therefore supports ML-guided virtual screening as an effective early-stage approach for expanding the search for structurally diverse CCR5 inhibitor candidates.

The subsequent virtual screening workflow further refined the ML-prioritized compounds according to their predicted drug-like, pharmacokinetic, safety, and target-binding properties. Of the 2,000 high-probability candidates, 170 satisfied the combined Lipinski, Ghose, Veber, and Egan criteria, supporting their favorable physicochemical characteristics and potential suitability for further drug development. ADMET screening subsequently reduced this set to 11 compounds with acceptable predicted pharmacokinetic and safety profiles. Computational ADMET assessment is increasingly important for early prioritization because experimental pharmacokinetic and toxicity testing is costly and time-consuming; recent deep-learning platforms such as Deep-PK have expanded such predictions across numerous ADMET endpoints [45]. Molecular docking showed that all 11 compounds interacted favorably with CCR5, with binding scores ranging from −7.1 to −11.4 kcal/mol. Notably, FDB020301 exhibited the most favorable predicted docking score (−11.4 kcal/mol), exceeding that of the reference CCR5 antagonist maraviroc (−10.5 kcal/mol), and interacted with key residues including TYR37, GLN280, GLU283, TRP86, and TYR108. These results support

FDB020301 as the leading computational candidate, although docking scores alone do not establish experimental binding affinity or inhibitory potency.

Molecular dynamics simulations provided additional support for the dynamic stability of the CCR5–FDB020301 complex beyond the static docking model. During the 10 ns simulation, the CCR5 backbone displayed an overall mean RMSD of 0.2884 ± 0.0414 nm, whereas FDB020301 showed a mean RMSD of 0.2105 ± 0.0470 nm, with only a transient ligand deviation around 4.65 ns followed by stabilization. The relatively stable RMSF profile, particularly around the ligand-binding region, indicated that ligand association did not cause substantial destabilization of the receptor. Similarly, the nearly constant Rg value of 2.5232 ± 0.0171 nm and the comparable average values during the final 5 ns suggested preservation of receptor compactness. Hydrogen-bond, SASA, PCA/FEL, and MM/GBSA analyses provided further evidence of favorable complex behavior, with a calculated MM/GBSA binding free energy of −35.19 ± 2.94 kcal/mol and substantial contributions from residues such as TRP86, TYR89, TYR108, and TYR37. Recent reviews recognize MD simulations as valuable tools for evaluating conformational dynamics, binding-pose stability, and lead optimization in drug discovery [46]. Collectively, these findings support the computational stability of FDB020301 within the CCR5 binding pocket and justify its further experimental validation.

Conclusion

As part of the present study, 11 lead candidates were identified as potential CCR5 antagonists through an integrated machine learning-driven *in silico* drug discovery approach. The XGBoost-based screening strategy enabled the identification of drug-like compounds with favorable predicted pharmacokinetic and safety profiles compared with maraviroc, the only FDA-approved CCR5 inhibitor. ADMET analysis indicated that the selected compounds exhibited desirable drug-like properties, including good absorption, high permeability, central nervous system penetration, and potential P-glycoprotein inhibitory activity, while showing no predicted mutagenic, carcinogenic, cardiotoxic, or other major toxicity risks. Molecular docking provided further support for their potential CCR5 inhibitory activity, with FDB020301 emerging as the top-ranked candidate and exhibiting a more favorable docking score (−11.4 kcal/mol) than maraviroc (−10.5 kcal/mol). FDB020301 also established favorable interactions with key CCR5 binding-site residues, including TYR37, GLN280, GLU283, TRP86, and TYR108. Molecular dynamics simulation further confirmed the stability of the CCR5–FDB020301 complex, as indicated by stable RMSD, RMSF, radius of gyration, hydrogen-bond, SASA, and PCA/FEL profiles, together with a favorable MM/GBSA binding free energy of −35.19 ± 2.94 kcal/mol. Therefore, the integrated application of machine learning, ADMET profiling, molecular docking, and molecular dynamics simulations identifies FDB020301 as a promising computationally validated CCR5 antagonist and potential lead molecule for further experimental evaluation and optimization toward the development of novel antiviral therapeutics.

Acknowledgements The authors have nothing to report

Author contributions A.S. and M.S conceived and designed the study. E.M interpreted the data, wrote the manuscript. All authors contributed to and have approved the final manuscript

Funding The present research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data Availability All data supporting the findings of this study are available within the paper and its Supplementary Information.

Ethics declaration: not applicable.

Declarations

Conflict of interest The authors declare no conflict of interest

References

1. UNAIDS (2026) Global HIV & AIDS statistics—Fact sheet. https://www.unaids.org/en/resources/fact-sheet. Accessed 2026.
2. Wang H, Yuan J, Wang H (2026) Research progress on HIV-1 structural proteins and antiviral therapies. Front Immunol 17:1805597. https://doi.org/10.3389/fimmu.2026.1805597
3. Elalouf A, Elalouf H, Maoz H (2025) HIV-1 entry mechanisms: protein-host receptor interactions and membrane fusion dynamics. Front Biosci (Landmark Ed) 30(10):37412. https://doi.org/10.31083/FBL37412
4. Yukselten Y, Wishah H, Li JA, Sutton RE (2025) Targeting CCR5: a central approach to HIV treatment and cure strategies. Virology 603:110375. https://doi.org/10.1016/j.virol.2024.110375
5. Goulart LM, Della Senta DO, Lourenço DA, Kremer FS (2026) Integrated virtual screening and molecular dynamics reveal novel Maraviroc analogues as optimized CCR5 entry inhibitors for HIV-1. J Mol Graph Model 148:109521. https://doi.org/10.1016/j.jmgm.2026.109521
6. Yin X, Xiao D, Yang J (2026) Emerging insights into CC and CXC chemokines and their receptors in Mycobacterium tuberculosis infection. FEBS Open Bio 16:1244–1258. https://doi.org/10.1002/2211-5463.70269
7. Niazi SK, Mariam Z (2024) Computer-aided drug design and drug discovery: a prospective analysis. Pharmaceuticals 17:22. https://doi.org/10.3390/ph17010022
8. El Sherbiny MM, Abdelfattah MG, Takieldeen AE, El-Din Mostafa H, Nafea HB (2026) An XGBoost-based predictive framework for diabetes mellitus multi-classification. Sci Rep 16(1):25519. https://doi.org/10.1038/s41598-026-65844-3
9. Altharawi A, Alqahtani SM (2026) Integrative computational chemistry approaches in modern drug discovery: advances in docking, pharmacophore modeling, molecular dynamics, and virtual screening. Pharmaceutics 18(5):565. https://doi.org/10.3390/pharmaceutics18050565
10. Zhao Y, He X, Wan Q (2024) Combined machine learning models, docking analysis, ADMET studies and molecular dynamics simulations for the design of novel FAK inhibitors against glioblastoma. BMC Chem 18:203. https://doi.org/10.1186/s13065-024-01316-x

11. Islam A, Jena D, Mondal NS, Teli A, Mondal S, Gautam MK (2025) In-silico approaches for drug designing technology: bridging discovery and development. Curr Drug Discov Technol 22. https://doi.org/10.2174/0115701638326869250207060616
12. Sanober MD, Mamidala E (2026) Apigenin derivatives as promising norovirus RdRp inhibitors: insights from in silico docking and molecular dynamics studies. Xenobiotica 56:121–131. https://doi.org/10.1080/00498254.2025.2566212
13. Alzunaidy NA (2026) Machine learning exploration of food-derived chemical space for potential nutritional metabolic regulators targeting dipeptidyl peptidase-4. Pharmaceuticals 19(3):349. https://doi.org/10.3390/ph19030349
14. Mei S (2024) Predicting health effects of food compounds via ensemble machine learning. Int J Food Sci Technol 59:2547–2557. https://doi.org/10.1111/ijfs.16992
15. Bozkır İ, İbişoğlu MS, Kayıkçıoğlu Bozkır İ, Güler Hİ (2026) Computational prioritization of multi-target inhibitors: explainable QSAR and docking-based discovery of dual AChE/BACE1 chemotypes. J Comput Aided Mol Des 40(1):54. https://doi.org/10.1007/s10822-025-00757-3
16. Wiens M, Verone-Boyle A, Henscheid N, Podichetty JT, Burton J (2025) A tutorial and use case example of the eXtreme gradient boosting (XGBoost) artificial intelligence algorithm for drug development applications. Clin Transl Sci 18. https://doi.org/10.1111/cts.70172
17. Liu T, Hwang L, Burley SK, Nitsche CI, Southan C, Walters WP, Gilson MK (2025) BindingDB in 2024: a FAIR knowledgebase of protein-small molecule binding data. Nucleic Acids Res 53(D1)–D1644. https://doi.org/10.1093/nar/gkae1075
18. Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, Li Q, Shoemaker BA, Thiessen PA, Yu B, Zaslavsky L, Zhang J, Bolton EE (2025) PubChem 2025 update. Nucleic Acids Res 53–D1525. https://doi.org/10.1093/nar/gkae1059
19. RDKit (2025) RDKit: open-source cheminformatics. Release 2025.09.1. https://www.rdkit.org. Accessed 2026.
20. Li J, Qu X, Zhang W, Zhong S (2026) Collision-free Morgan fingerprints: a principled approach to enhance machine learning performance and interpretability in chemistry. J Cheminform 18(1):46. https://doi.org/10.1186/s13321-026-01170-0
21. 21. Python Software Foundation (2024) Python language reference, version 3.12.3. https://www.python.org. Accessed 2026.
22. 22. Anaconda Inc (2026) Anaconda Distribution, version 2026.10. https://www.anaconda.com. Accessed 2026.
23. 23. Spyder Project (2026) Spyder: the scientific Python development environment. https://www.spyder-ide.org. Accessed 2026.
24. Kaya AO (2026) Interpretable machine learning-driven QSAR modeling for coagulation factor X inhibitors: from molecular descriptors to predictive potency. J Comput Aided Mol Des 40:53. https://doi.org/10.1007/s10822-025-00758-2
25. Cabot JH, Ross EG (2023) Evaluating prediction model performance. Surgery 174(3):723–726. https://doi.org/10.1016/j.surg.2023.05.023
26. FooDB (2026) The Food Database. University of Alberta, Canada. Accessed 2026.

27. Reese TC, Devineni A, Smith T, Lalami I, Ahn JM, Raj GV (2024) Evaluating physicochemical properties of FDA-approved orally administered drugs. Expert Opin Drug Discov 19(2):225–238. https://doi.org/10.1080/17460441.2023.2275617
28. Patra S, Paul A, Shand H, Ghosal S, Ghorai S (2025) In silico identification of anticancer flavonoids as dengue virus replication inhibitors: a molecular docking and simulation approach. J Mol Model 31(11):287. https://doi.org/10.1007/s00894-025-06516-3
29. Caminero Gomes Soares A, Marques Sousa GH, Calil RL, Goulart Trossini GH (2023) Absorption matters: a closer look at popular oral bioavailability rules for drug approvals. Mol Inform 42(11). https://doi.org/10.1002/minf.202300115
30. Daina A, Michielin O, Zoete V (2017) SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. Sci Rep 7:42717. https://doi.org/10.1038/srep42717
31. Myung Y, de Sá AGC, Ascher DB (2024) Deep-PK: deep learning for small molecule pharmacokinetic and toxicity prediction. Nucleic Acids Res 52–W475. https://doi.org/10.1093/nar/gkae254
32. Swapna K, Srujana M, Mamidala E (2024) Identification of steroidal cardenolides from Calotropis procera as novel HIV-1 PR inhibitors: a molecular docking & molecular dynamics simulation study. Indian J Med Res 160:78–86. https://doi.org/10.25259/IJMR_2115_23
33. Abdelwahab AA, Elattar MA, Fawzi SA (2025) Advancing ADMET prediction for major CYP450 isoforms: graph-based models, limitations, and future directions. Biomed Eng Online 24(1):93. https://doi.org/10.1186/s12938-025-01412-6
34. Marković V, Szczepańska A, Berlicki Ł (2024) Antiviral protein-protein interaction inhibitors. J Med Chem 67(5):3205–3231. https://doi.org/10.1021/acs.jmedchem.3c01543
35. BIOVIA, Dassault Systèmes (2024) BIOVIA Discovery Studio Visualizer, Release 2024 [Computer software]. Dassault Systèmes, San Diego, CA, USA.
36. Roney M, Mohd Hisam NS, Uddin M et al (2026) Repurposing of natural products for spinocerebellar ataxia type 3 using integrated network pharmacology and in silico approaches. Sci Rep 16:7332. https://doi.org/10.1038/s41598-025-30652-8
37. Rajput A, Bhamare KT, Thakur A, Kumar M (2022) Biofilm-i: a platform for predicting biofilm inhibitors using quantitative structure–relationship (QSAR) based regression models to curb antibiotic resistance. Molecules 27(15):4861. https://doi.org/10.3390/molecules27154861
38. Yu D, Li H, Liu Y, Yang X, Yang W, Fu Y, Zuo YA, Huang X (2024) Application of the molecular dynamics simulation GROMACS in food science. Food Res Int 190:114653. https://doi.org/10.1016/j.foodres.2024.114653
39. Sathish Kumar A, Mamidala E (2026) Pharmacophore-based identification and molecular docking of CCR5 inhibitors: insights from dynamic simulations. J Recept Signal Transduct Res 46:113-121. https://doi.org/10.1080/10799893.2026.2701217
40. Gorelov S, Titov A, Tolicheva O, Konevega A, Shvetsov A (2024) Determination of hydrogen bonds in GROMACS: a new implementation to overcome memory limitation. J Chem Inf Model 64(16):6241–6246. https://doi.org/10.1021/acs.jcim.3c02087

41. K D, Venugopal S (2024) Molecular docking and molecular dynamic simulation studies to identify potential terpenes against Internalin A protein of Listeria monocytogenes. Front Bioinform 4:1463750. https://doi.org/10.3389/fbinf.2024.1463750
42. Zhao SS, Wang YJ, Tang L, Guo B, Wang L, Zhang JQ, Yang SG (2023) Identifying novel selective PPO inhibitors through structure-based virtual screening and bio-evaluation. RSC Adv 13:10873-10883. https://doi.org/10.1039/D2RA08006K
43. Rainio O, Teuho J, Klén R (2024) Evaluation metrics and statistical tests for machine learning. Sci Rep 14:6086. https://doi.org/10.1038/s41598-024-56706-x
44. Tarwidi D, Pudjaprasetya SR, Adytia D, Apri M (2023) An optimized XGBoost-based machine learning method for predicting wave run-up on a sloping beach. MethodsX 10:102119. https://doi.org/10.1016/j.mex.2023.102119
45. Yang C, Gong C, Zhang Z, Fang J, Li W, Liu G, Tang Y (2025) *In silico* prediction of pKa values using explainable deep learning methods. J Pharm Anal 15:101174. https://doi.org/10.1016/j.jpha.2024.101174
46. Yan D, Ma Y, Chen X, Deng S, Wang Q (2025) Molecular dynamics-driven drug discovery. Phys Chem Chem Phys 27:12633-12651. https://doi.org/10.1039/D5CP00380F

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTableS1FoodDBdruglikenessrulepassed.csv