scientific reports

OPEN

Integrating machine learning and structure-based approaches for repurposing potent tyrosine protein kinase Src inhibitors to treat inflammatory disorders

Muhammad Waleed Iqbal[1], Muhammad Shahab[1], Zakir ullah[1], Guojun Zheng[1], Irfan Anjum[2], Gamal A. Shazly[3], Atrsaw Asrat Mengistie[4]✉, Xinxiao Sun[1]✉ & Qipeng Yuan[1]✉

Tyrosine-protein kinase Src plays a key role in cell proliferation and growth under favorable conditions, but its overexpression and genetic mutations can lead to the progression of various inflammatory diseases. Due to the specificity and selectivity problems of previously discovered inhibitors like dasatinib and bosutinib, we employed an integrated machine learning and structure-based drug repurposing strategy to find novel, targeted, and non-toxic Src kinase inhibitors. Different machine learning models including random forest (RF), k-nearest neighbors (K-NN), decision tree, and support vector machine (SVM), were trained using already available bioactivity data of Src kinase targeting compounds. The performance evaluation of these models demonstrated SVM as the best model, which was further utilized to shortlist 51 highly potent compounds by screening an FDA-approved library of 1040 drugs. Molecular docking and molecular dynamic simulation were subsequently employed to evaluate the binding affinity and stability of the proposed compounds. Orlistat, acarbose and afatinib were identified as the potent leads, demonstrating stable conformations and stronger interactions, validated by root mean square deviation (RMSD), root mean square fluctuation (RMSF), radius of gyration (RoG), and hydrogen bond analyses. Molecular Mechanics/Generalized Born Surface Area (MMGBSA) analysis validated their binding affinities by providing comparably lower binding free energies for orlistat (− 33.4743 ± 3.8908), acarbose (− 19.5455 ± 5.4702), and afatinib (− 36.4944 ± 5.4929) than the control, dasatinib (− 13.7785 ± 5.8058). Finally, toxicity analysis revealed orlistat and acarbose as the possible safer therapeutics by eliminating afatinib as it showed significant toxicity concerns. Our investigation supports the advance computational methods utilization in the field of drug discovery and suggest further experimental validation of proposed inhibitors of Src kinase for their safer use against inflammatory diseases. The ultimate aim of this study is to advance the development of effective treatments for inflammatory diseases, linked with Src overexpression.

Keywords Src kinase, Inflammatory diseases, Machine learning, Docking, Molecular dynamic simulation

Immune-mediated inflammatory diseases are the most prevalent group of conditions, which are characterized by dysregulated immune responses[1]. These chronic conditions encompasses a variety of diseases, including rheumatoid arthritis[2], systemic lupus erythematosus (SLE)[3], atherosclerosis[4], inflammatory bowel disease (IBD)[5], psoriasis[6], osteoporosis[7], and pemphigoid diseases[8]. The prevalence of these inflammatory diseases varies among them as rheumatoid arthritis has been reported in most of the cases, which are estimated to be 1% of all the adults worldwide[9]. On the other hand, Sjögren's syndrome, one of the inflammatory disease, has been investigated to show potential in least of the world population (i.e. about 0.1–0.6%)[10]. Among them, systemic

[1]State Key Laboratory of Chemical Resources Engineering, Beijing University of Chemical Technology, Beijing 100029, People's Republic of China. [2]Department of Basic Medical Sciences, Shifa College of Pharmaceutical Sciences, Shifa Tameer-e-Millat University, Islamabad 44000, Pakistan. [3]Department of Pharmaceutics, College of Pharmacy, King Saud University, Riyadh 11451, Saudi Arabia. [4]Department of Biology, Bahir Dar University, P.O. Box 79, Bahir Dar, Ethiopia. ✉email: smartreservcher@gmail.com; sunxx@mail.buct.edu.cn; yuanqp@mail.buct.edu.cn

lupus erythematosus (SLE) has been reported as the leading cause of deaths in the older studies, suggesting 34% of deaths annually[11]. Major causes of deaths in inflammatory conditions are mostly cardiovascular complications[12], infections[13], organ damage[14] and sometimes cancer[15].

Genetic mutations in various genes have been considered as a key factor, involved in multiple inflammatory immune diseases[16]. These genetic mutations often result in the overexpression of proteins, encoded by relevant genes. The overexpression of certain proteins, especially protein kinases, may contribute to the abnormal growth of cells or disruption of normal cellular homeostasis and thus significantly play role in inflammation[17]. One of the most important gene in this regard is Src kinase, which encodes a non-receptor tyrosine kinase protein. In normal conditions, it plays a crucial role in various cellular processes including proliferation, migration, differentiation, survival and immune responses[18]. It primarily contains three domains including SH1 domain, SH2 domain and SH3 domain. All of these domains involved in performing diverse functions as SH1 domain (also known as kinase domain) is responsible for catalytic activity by phosphorylating tyrosine residues on substrate proteins, SH2 domain binds to the resulted phosphorylated tyrosine residues, and SH3 domain binds with proline-rich sequences to facilitate interactions with other proteins[19]. Genetic mutations and resulted overexpression often influence its normal functioning, leading to either loss or gain of function, which result in inactive kinase or enhanced kinase activity respectively[20]. In both the conditions, the inflammation is promoted as inactive kinase potentially leads to the impaired cell growth whereas overactive kinase continuously signal without proper regulations. For example, tyrosine kinase protein Src kinase plays a pivotal role in T-cell and B-cell receptor signal transduction. Mutations in Src kinase enhance its activity, leading to the hyper activation of B-cells and T-cells. These alterations in receptor signal transduction resulted in increased protein phosphorylation, leading to a fatal inflammatory condition known as systemic lupus erythematosus (SLE)[21,22]. Many studies have reported Src kinase involvement in immune cell signaling pathways including immunoreceptor, integrin, and c-type lectin signaling, which potentially contribute to the inflammatory processes related to atherosclerosis. Moreover, over activation of Src kinase in smooth muscle cells may also stimulates their migration from media to the intima of the artery wall, forming atherosclerotic plaques[23]. A study also reported PKCθ, which is a Src kinase family kinase, plays a crucial role in T-cell mediated inflammation in inflammatory bowel disease (IBD)[24]. Some studies also investigated potential inflammatory role of Src kinase in various inflammatory pemphigoid diseases[25]. It has been reported that tyrosine protein kinase Src kinase plays a vital role in signaling through Fc receptors, which involve in binding and recognizing of immune complexes at the dermal-epidermal junction in pemphigoid diseases.

As inflammatory conditions present key health issues, there already have been some inhibitors designed for a variety of proteins, reported to involve in inflammatory processes. Likewise, there have been some researches done to find a potential inhibitor against Src kinase activity. Significant developments have been reported, involving the use computational approaches for finding the potential inhibitors of Src in drug discovery field. For instance, a study utilized unbiased molecular dynamic simulation to check the binding process of dasatinib, an FDA-approved anti-cancer drug, with Src kinase to inhibit its activity[26]. Likewise, a recent study incorporated computational simulation and 3D-QSAR (quantitative structure-activity relationship) strategies to propose a Src inhibitor named CID_70144047[27]. Apart from all the identified inhibitors, dasatinib have been demonstrated to show the highest inhibitory effect as a second-generation Src tyrosine kinase inhibitor (STKI), which has also been approved from FDA[28]. Furthermore, Bosutinib is another FDA-approved drug that has been found to show the high inhibitory effect against Src kinase[29]. Despite their potential against Src kinase, these inhibitors showed some limitations like limited efficacy, potential toxicity, resistance development and lack of specificity[30]. All these challenges encouraged us to find a better Src kinase inhibitor, which can have ability to overcome these limitations. In this study, a consensus approach integrating machine learning with structure-based repurposing of freely accessible FDA-approved drug library was applied. The most accurate and efficiently trained machine learning model was employed to check the bioactivity of complete drug library. Subsequently, the MOE software was employed for molecular docking of the highly reactive bioactive drugs with Src Kinase, due to its proven efficacy in accurately modeling protein-ligand interactions, as demonstrated in numerous studies[31,32]. Molecular dynamic simulation strategy further validated the atomic level stability of our proposed drugs, followed by a toxicity analysis. This approach could revolutionize the drug repurposing process by significantly reducing the time and cost associated with traditional drug development methods. Furthermore, this approach enables for further preclinical and clinical testing of our proposed drugs and also highlights the comprehensive use of computer-aided strategies in the field of drug designing.

Methodology

To find a potential inhibitor against Src kinase activity, a machine learning-based approach was integrated with the structure-based repurposing. By applying a most accurately-trained machine learning model, an already available FDA-approved library was screened, containing 1040 drugs. Comprehensive computational analysis including molecular docking, molecular dynamic simulation and post-trajectory analysis like root mean square deviation (RMSD), root mean square fluctuation (RMSF) radius of gyration (RoG), hydrogen bond analysis, and MMGBSA/MMPBSA analysis allowed us to delve deeper into the identification of a novel Src kinase inhibitor. The complete workflow of our used approach is illustrated in Fig. 1.

Retrieval and preprocessing of bioactive data

The complete dataset, consisting 3570 bioactive compounds targeting tyrosine kinase protein Src kinase, was acquired from ChEMBL database (https://www.ebi.ac.uk/chembl/) by utilizing ChEMBL websource client (Uniprot ID: P12931)[33]. Subsequently, an open source and widely-used toolkit for chemoinformatics purposes, RDKit, was employed to preprocess and filter bioactive compounds on the basis of their issues with canonical smiles and possible duplicates[34]. Following this, the IC_{50} (half maximal inhibitory concentration) values were

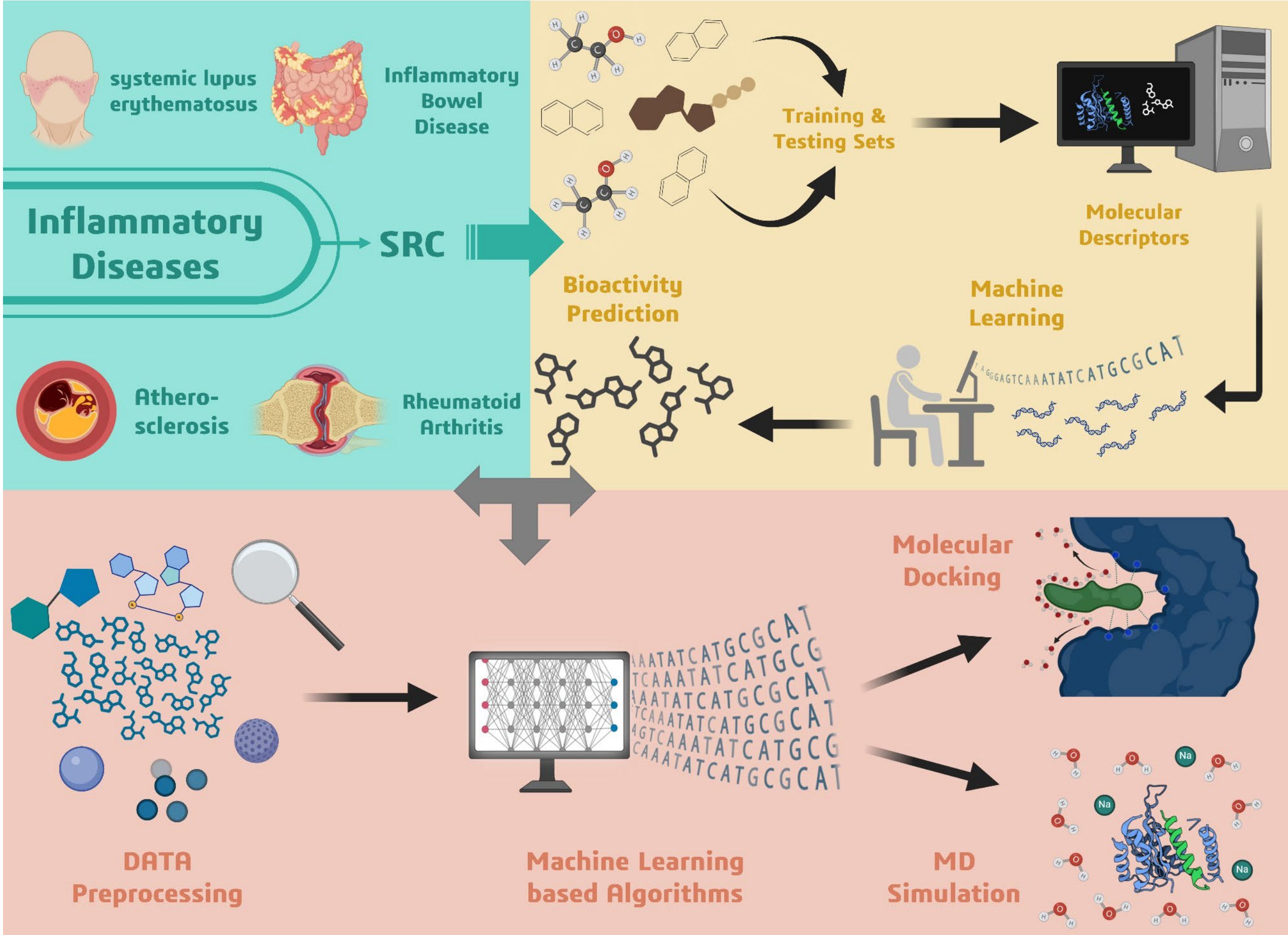

Fig. 1. A process diagram representing the overall workflow of the methodology employed in this study to find potent inhibitors of Src kinase.

extracted from the filtered data and dataset was distributed into three categories based on their IC_{50} values[35]. The compounds with IC_{50} values lesser than 1000 nM (i.e. IC_{50} <= 1000 nM) were categorized as active compounds and the compounds with IC_{50} values greater than 10,000 nM (IC_{50} => 10000 nM) were classified as inactive compounds. The compounds, having IC_{50} values between 1000 nM and 10,000 nM were distributed as intermediate compounds. IC_{50} values of the filtered dataset were then converted into pIC_{50} values using math package, to understand the detailed potency of bioactive molecules[36].

Comprehensive filtration using ADMET and PAINS analyses

After having our bioactive dataset in well-classified form, a comprehensive filtration of this dataset was conducted by employing globally-accepted strategies. Lipinski's descriptors and pandas tools were utilized from RDkit package to identify compounds, which fulfil the Lipinski's rule of 5 (Ro5). According to this rule, a compound has a drug-like properties if its molecular weight is less than 500 Daltons, can accept a maximum of 10 hydrogen bonds, can donate a maximum of 5 hydrogen bonds, and its KOW (n-octanol/water partition coefficient) is less than 5[37]. To identify Ro5 fulfilling bioactive compounds, Lipinski's descriptor were assigned to whole of the dataset. After having Ro5 fulfilled compounds, intermediate and inactive categorized compounds were excluded as a filtration step. Subsequently, seaborn and Matplotlib packages were utilized to visualize the graphical interface of Ro5 fulfilled compounds[38]. Furthermore, a Pan Assay Interference Compounds (PAINS) algorithm was applied using RDKit to further validate the Ro5 fulfilled compounds and to exclude the possible false positives[39].

Descriptor generation and final dataset preparation

Following ADMET and PAINS analyses, further filtration was applied on already filtered dataset before its final preparation. This is because sometimes bioactive dataset contains such compounds which exhibits high toxicity and mutagenesis profiles due to their unfavorable pharmokinetic properties. Filtering out these sort of compounds is always essential for more efficient and precise dataset generation[40]. For this purpose, a list of already defined substructures, by the research group of Brenk et al.[41], was employed using RDkit library. This list consisted of substructures including phosphates and sulfates (unfavorable pharmacokinetic properties), nitro groups (mutagenic), and thiols (highly reactive), which were compiled to screen compounds linked with various diseases. After having our comprehensively filtered dataset, PaDEL (Prediction of Activity Spectra for Substances) descriptors were utilized to generate molecular fingerprinting properties of the respective bioactive compounds in binary form[42]. Moreover, machine was further assisted by Pandas library[43] to shortlist the relevant descriptors and reduce dimensionality from described dataset.

Multiple machine learning models building

After having well-prepared dataset of comprehensively filtered bioactive compounds, it was utilized to train diverse machine learning models including Decision Tree, Random Forest, K-Nearest Neighbors, and Support Vector Machine. All of these machine learning algorithms are widely utilized for drug designing and data analysis purposes as each of these models has its distinct classification and regression characteristics. Decision Tree consists of a tree-like structure containing a group of nodes where each node perform classification and regression tasks by splitting input dataset into output values[44], whereas Random Forest involves various decision trees to train and test the input data[45]. Likewise, K-Nearest Neighbors (K-NN) is also an instance-based machine learning algorithm that considers its nearest neighbors to classify the input dataset[46]. Support Vector Machine is also widely utilized to perform both supervised and unsupervised tasks on the given input dataset but it predominately best at performing supervised learning tasks[47]. Before delving deeper with those machine learning models, the bioactive dataset was assigned into 80/20 fashion, in which 80% data has to be trained and 20% data to test[48]. Low variance features of the bioactive dataset were removed using variance threshold function of Scikit-learn package to train the ML models more accurately. Following this, cross-correlation matrix analysis was performed on remaining features by using corr () function from pandas library to identify and eliminate non-correlative features[49]. Furthermore, recursive feature elimination (RFE), a strategy to select more relevant features recursively, was applied on binary dataset using Scikit-learn package[50]. After reducing overfitting features and computational complexity using RFE function, the bioactive dataset was trained and tested. Finally, a permutation importance analysis was conducted by employing permutation_importance library from scikit-learn, to identify the corrected features of best-trained model that are involved in final prediction[51].

Cross-validation of trained ML models

After training various machine learning algorithms based on Src kinase targeting bioactive data, LazyPredict library and Scikit-learn package were employed to evaluate and validate those models' performance[52]. Initially, a LazyRegressor from LazyPredict library was employed to automate the training and evaluation of multiple regression models including Random Forest (RF), K-Nearest Neighbors (K-NN), Support Vector Machine (SVM) and Decision Tree, which were compared with the other 37 machine learning models. The LazyRegressor fitted the models on the training data and produced performance metrics for each model based on Root mean square error (RMSE), R-squared (R^2) values, and overall computation time. The RMSE is a widely used technique to evaluate and compare the accuracy of various regression models[53], whereas R-squared values measure the proportion of dependent variables of regression models, with respect to their independent variables[54]. The performance metrics obtained from LazyRegressor were used to compare the models comprehensively. To graphically represent the performance comparison of trained machine learning models with the diverse models, seaborn and Matplotlib packages were employed[38].

Potential hits selection and docking

Following the performance evaluation of the trained machine learning models, the best model was employed to deal with our targeted drugs. This model was applied on already available FDA-approved library of 1040 drugs to find potent Src kinase inhibitors from them. When applying this model on our drug library, bioactivities of all the dealt drugs in the form of pIC50 values were given as output. The top drugs, having higher pIC50 values, were selected and screened through a widely used computational technique for ligand-receptor binding evaluation, known as molecular docking[55]. For this purpose, the PDB structure of tyrosine protein kinase Src kinase (PDB ID: **1FMK**) was acquired from protein data bank (PDB) (https://www.rcsb.org/)[56]. The structure was prepared prior to docking, by removing ligands and water molecules using Molecular Operating Environment (MOE) software[57,58]. 3D protonation was done on the amino acids of target protein at psychological pH 7.4, which is commonly used in computational studies involving biological systems[59]. MOE software automatically assigned the protonation states to amino acids based on this pH. Subsequently, energy minimization was also conducted to remove abnormalities in the structure. By employing three different scoring functions (Affinity dG, London dG, GBVI/WSA), the in-built triangle matcher placing method of MOE was utilized to analyze the interaction profile of Src kinase residues[60,61]. While screening the drug library, two key parameters (i.e. RMSD and binding affinity) were considered. The top drugs, having binding affinity less than −7.5 and RMSD values less than 2, were shortlisted for the comprehensive molecular dynamic simulation analysis[62]. Discovery studio[63] and PyMol[64] visualization software were used to examine the binding interactions of docked drugs.

Molecular dynamic simulation of top hits

For the atomic-level stability validation of potent leads from molecular docking and to check their individual atoms mobility in a well prepared environment, molecular dynamic simulation[65] was employed. Initially, ff19SB force field and Amber22 package were utilized for setting up the system[66]. Antechamber and parmchk2 packages of amber tools were employed to assign required parameters for all the drugs[67]. After having our prepared drugs, a complex preparation amber tool, tleap, was used to facilitate and construct the system. To neutralize the system before simulation, respective Na^+ and Cl^- ions were added in it[68]. Subsequently, the complex was prepared by solvating the system. To initiate the MD simulation, the system was minimized for removing any structural deformities. The energy minimization process was carried within two steps as the first step consisted of overall 2500 steps, including 1500 steps conjugate gradient and 1000 steps of steepest descent with constraints, whereas in the second step, the 1500 steps were assigned for conjugate gradient and 1000 steps for steepest descent but without constraints to rectify conflicts and collisions in the system[69]. Following the two-step energy minimization process, the system was heated for 50ps at about 300 K, in which Langevin thermostat algorithm[70] was utilized to control temperature and Berendsen barostat algorithms[71] was used to check overall system temperature. Furthermore, Amber22 Shake algorithm was applied to further strengthen covalent bonds' interactions[72]. Following this, the system was equilibrated for 1 ns as the pre-final step[73]. Finally, MD simulation of 300 ns for each complex was carried out using pmemd.cuda of amber22[74]. Finally, the resulted trajectories were analyzed using CPPTRAJ package[75] for comprehensive assessment.

Free energy calculations

Following the MD simulation of all shortlisted complexes, their binding free energies were evaluated using MMGBSA (Molecular Mechanics/Generalized Born Surface Area) and MMPBSA (Molecular Mechanics/Poisson-Boltzmann Surface Area) methods[76,77]. Both of these widely-used methods were employed to calculate the free energies of all the complexes over the entire production phases of 300 ns of simulation process. To delve deeper into the diverse conformational specs of simulated complexes, the parameters was set as overall 30,000 frames were captured at each interval of 1 ns (i.e. 100 frames). Binding free calculations were denoted by ΔG_{Bind}, calculated by following equations:

$$\Delta\ \mathrm{G}_{\mathrm{Bind}} - {}^{\mathrm{PB}}\!/_{\mathrm{GB}} = \Delta\ \mathrm{E}_{\mathrm{MM}} + \Delta\ \mathrm{G}_{\mathrm{SOL}\left(\frac{\mathrm{PB}}{\mathrm{GB}}\right)} - \mathrm{T}\Delta\ \mathrm{S}$$

$$\Delta\ \mathrm{E}_{\mathrm{MM}} = \Delta\ E_{internal} + \Delta\ E_{electrostatic} + \Delta\ E_{vdw}$$

$$\Delta\ \mathrm{G}_{\mathrm{SOL}\left(\frac{\mathrm{PB}}{\mathrm{GB}}\right)} = \Delta\ \mathrm{G}_{\mathrm{PB/GB}} + \Delta\ \mathrm{G}_{\mathrm{SA(PB/GB)}}$$

Where ΔE_{MM} is total gas phase energy (sum of $\Delta E_{internal} + \Delta E_{electrostatic} + \Delta E_{vdw}$). $\Delta G_{SOL(PB/GB)}$ is sum of polar and nonpolar addition to solvation estimated by PB or GB. The conformational entropy during binding is denoted by TΔS. The internal energy of the MM force field, $\Delta E_{internal}$, is obtained from bond, angle, and dihedral parameters.

Toxicity analysis

After comprehensively validating the proposed compounds as inhibitors of Src kinase, a final toxicity analysis was performed to ensure their safety inside the body. For this purpose, an open-source deep learning framework specifically designed for molecular property prediction, named Chemprop 1.5.2, was utilized that employs the graph neural networks (GNNs) to model molecular as graph[78]. Subsequently, Moleculenet (https://moleculenet.org/), a well-established benchmark platform, was chosen to retrieve high-quality dataset for molecular property prediction[79]. The chemprop model was trained based on the acquired dataset, and applied on our proposed drugs and control. The expected toxicity of proposed drugs, along with the control, was predicted and illustrated using matplotlib and seaborn packages[38].

Results

Src kinase is a noon-receptor protein kinase, which usually involves in a variety of cellular processes including proliferation, differentiation and migration but when it is overexpressed or being mutated, it can result in the progression of various inflammatory immune diseases. To inhibit its activity under abnormal conditions, we utilized machine learning and structure-based drug repurposing of FDA-approved drug library, which was further validation by molecular dynamic simulation and MMGBSA/MMPBSA approaches.

Retrieval, preprocessing and filtration of bioactive dataset

Tyrosine kinase protein Src kinase targeting 3570 bioactive compounds was acquired from ChEMBL database. From the complete acquired dataset, 743 compounds having issues with their smiles and contain duplicates were removed for further processing. All of the 2827 filtered compounds were, then, categorized as intermediate, active and inactive compounds to standardize their distribution. Their IC_{50} values were converted to a more efficient unit for potency, pIC_{50}. Matplotlib package was utilized to graphically visualize the bioactive dataset distribution (Fig. 2).

After converting bioactive dataset in pIC50 values, we performed further filtration by dealing it with the Lipinski's descriptors using Rdkit package. Out of 2827 bioactive compounds, only 1239 fulfilled Lipinski's rule of five based on their fingerprinting data. We, then, filtered the intermediate class as well from our remaining bioactive compounds, resulted in the exclusion of 297 compounds. Furthermore, PAINS filtration was applied on the remaining 942 bioactive compounds to check false positives among them, which resulted in the further elimination of 67 compounds. The remaining high quality bioactive compounds, fulfilling both ADMET and PAINS parameters, were then further processed. Graphical visualization of filtered compounds is presented in Fig. 3.

Substructures filtration and molecular descriptors generation

Further filtration to remove unwanted substructure from bioactive dataset was applied by defining an already available list of substructures. Out of 875 bioactive compounds, only 586 compounds were passed from this filtration test, eliminating 365 compounds as undesired substructures. The filtered compounds found to contain substructures including 79 aliphatic long chains, 64 Michael acceptors, 59 phosphor and 32 oxygen-nitrogen single bonds. These filtered 586 compounds were considered as the most accurate and reliable compounds to train machine learning models. The PaDEL descriptors were, finally, applied on these filtered compounds, which

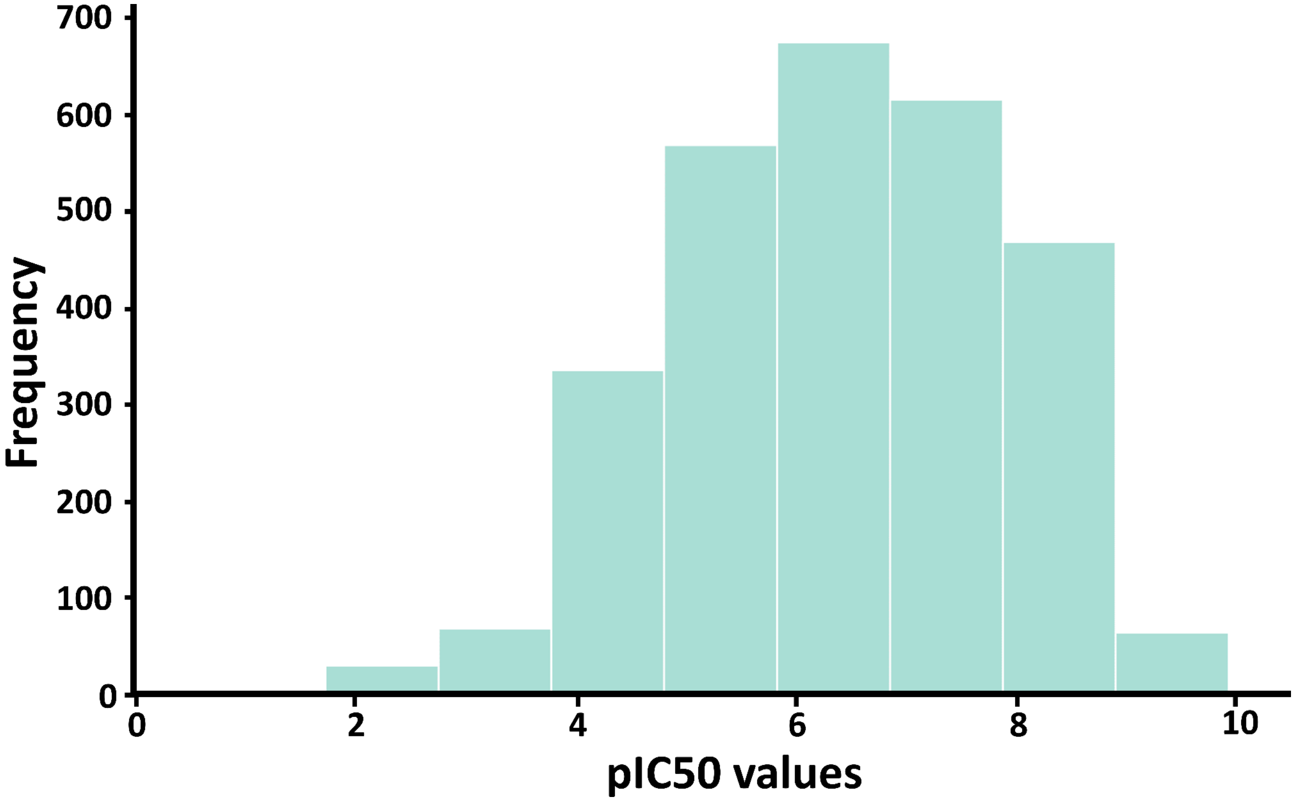

Fig. 2. Representing the varied pIC50 values of all the Src-targeted compounds, retrieved from ChEMBL.

assigned the local features of all the molecules in machine-recognizing binary form. These local features are the building blocks of each respective molecule, which form the basis of its pharmacological characteristics[80].

Model evaluation and performance validation

Following the generation of unique fingerprints for all the comprehensively filtered 586 molecules, we utilized them to train different machine learning algorithms including Random Forest (RF), K-Nearest Neighbor (K-NN), Support Vector Machine (SVM) and Decision Tree. Before training these models, we filtered out low variance features from fingerprinting data of those molecules. After having 190 filtered fingerprints out of 881 for each compound, we further applied cross-correlation matrix analysis, resulted in the removal of 98 features, as depicted in Fig. 4A. The final 92 refined features were subsequently validated using recursive feature elimination (RFE) method. RFE further filtered out 29 undesired fingerprints from each bioactive compound (Fig. 4B). After having the high quality fingerprinting data, we trained all of the four machine learning models in 80/20 fashion. From all of the four models, SVM model showed best performance as its R^2 (coefficient of determination) score was about 0.38, followed by random forest (0.30), k-nearest neighbors (0.25) and decision tree (– 0.35) (Fig. 4C). The permutation importance analysis in Fig. 4D represents the best features of SVM model, which were involved in prediction. The top five features having highest importance scores, including PubchemFP37 (0.033), PubchemFP391 (0.029), PubchemFP392 (0.022), PubchemFP758 (0.018), and PubchemFP261 (0.016), played the crucial role in the final prediction.

The performance of each trained machine learning model was evaluated by comparing them with 37 other models' performance. Here, SVM model demonstrated the best performance among all, as indicated by RMSE and R-squared values. The R-squared value (Fig. 5A) for SVM model was about 0.4, which was highest among all other models including random forest (0.30), K-NN (0.25), and decision tree (– 0.35). Likewise, RMSE values (Fig. 5B) for SVM model was lowest among all as it was near to 0, indicating a highly accurate and efficient performance. Moreover, computational time taken by SVM model for making prediction was about 0.1 s, which suggested it as a very efficient model. While external validation sets were not used in this study, comprehensive model evaluation by LazyRegressor was deemed sufficient to validate the results. The overall best accuracy and efficiency of SVM model encouraged us to further apply it on our drug library.

Potential hits selection and docking

SVM model was emerged as the best model after performance validation with all the other machine learning models. We, then, employed it to screen our desired drugs on the basis of trained bioactive dataset. We applied the trained model on FDA-approved library of 1040 drugs to identify the best hits based on their pIC50 values. The model gave varied pIC50 values from 4 to 7.7 as output. Many studies reported pIC50 values between 6 and 8 indicates moderately active drugs, showing good potency[81,82]. Some studies also suggest that the higher pIC50

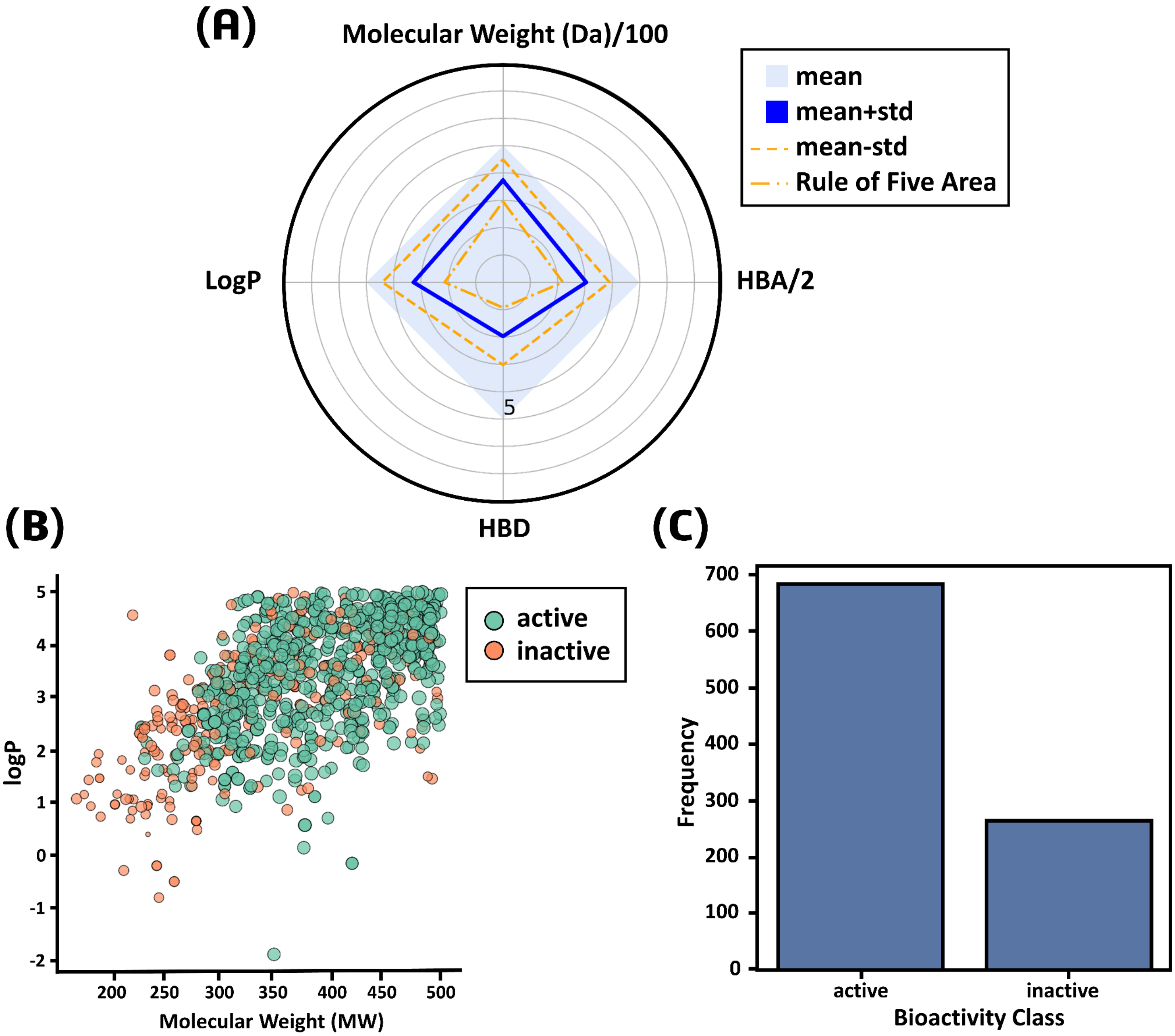

Fig. 3. **(A)** A radar plot representing the deviation of Ro5-fulfilled bioactive compounds from their mean **(B)** A scatter plot representing the ratio of logP and molecular weight of bioactive compounds in both active and inactive categories **(C)** A bar plot representing the number of filtered compounds from active and inactive class.

values demonstrates higher inhibitory effect of the drug[83–85]. Therefore, we selected top 51 compounds, having pIC50 values over 6.5, for further screening through molecular docking. For initiating docking process, energy minimization and 3D protonation of Src kinase structure was done using MOE as a preprocessing docking step. Src kinase bound ligands were removed and polar hydrogen were added to get the good binding affinity between drugs and receptor. After start of the docking, 41 drugs were screened simultaneously through prepared Src kinase 3D structure, resulting in the varied binding affinities ranging from − 4 to − 8.7 kJ/mol. Among all the docked compounds, top 3 compounds including orlistat (S-score − 8.7, RMSD 1.6), acarbose (S-score − 7.9, RMSD 1.9) and Afatinib (S-score − 7.7, RMSD 1.6) were chosen as they fulfilled previously indicated threshold of S-Score below − 7 and RMSD below 2. This criteria was set based on previous molecular docking studies involving kinase inhibitors, where docking scores below − 6 to − 7 kJ/mol were shown to indicate strong binding affinity and meaningful molecular interactions[86]. For instance, previous studies on ATP-competitive inhibitors have shown that drugs in this range of docking scores have high bioactivity in clinical trials. This guarantees that only inhibitors with the potential to be effective were identified for additional research[87]. Furthermore, an already FDA-approved Src kinase drug named dasatinib was also taken as control to evaluate and compare its activity with our proposed drugs. Docking results of dasatinib with receptor Src kinase showed its S-score (-6.8) and RMSD (2.52) were comparatively low than our proposed drugs. Docking results of top hits, along with the control drug, with receptor Src kinase have been shown in Table 1.

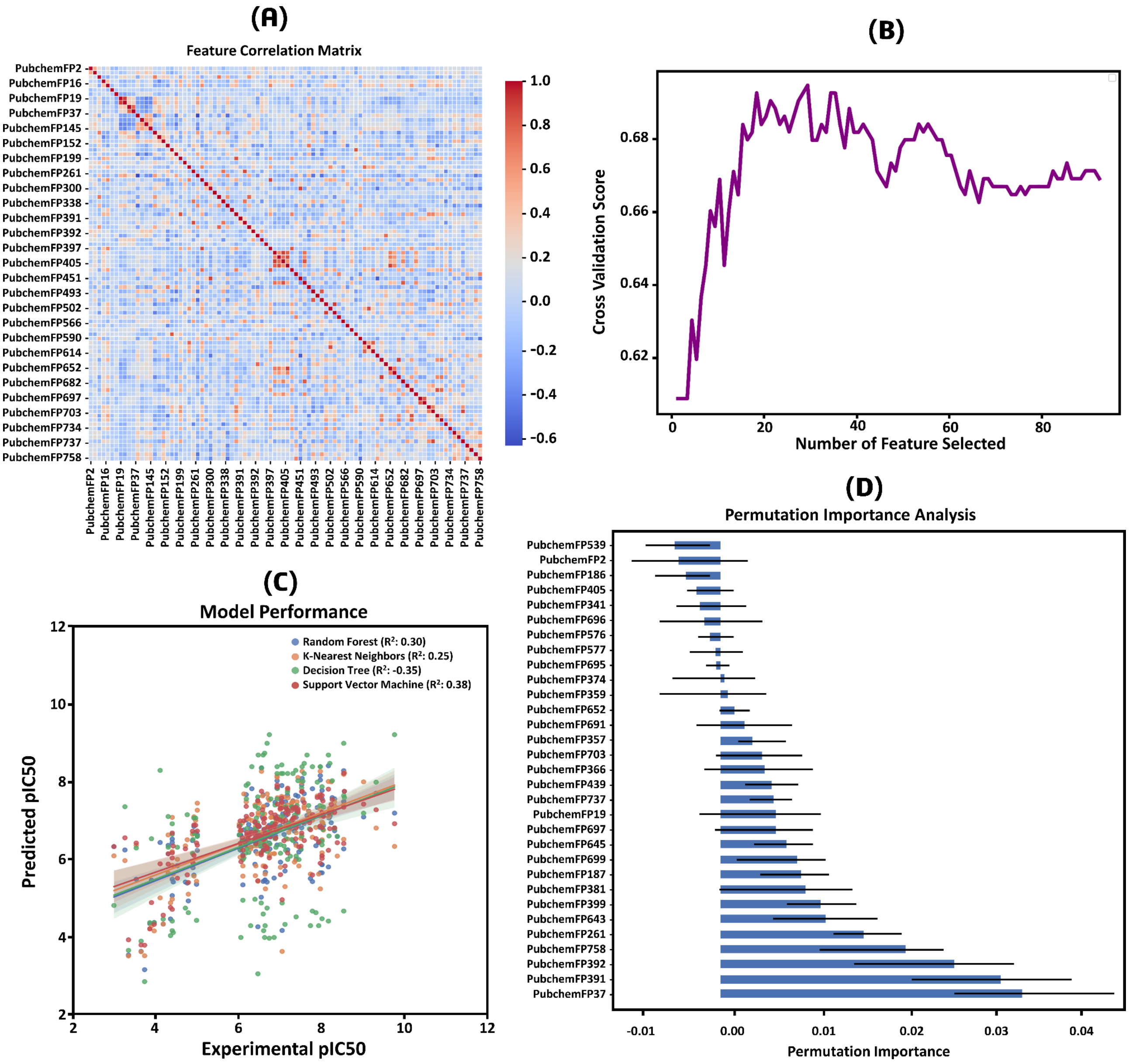

Fig. 4. **(A)** A heatmap representing the correlated features of refined bioactive compounds **(B)** A line graph representing the performance of each fingerprinting feature, validated by recursive function elimination (RFE) analysis **(C)** A scatter plot comparing existing experimental pIC_{50} values with predicted pIC_{50} values from different trained machine learning models **(D)** A graph representing the significance of each refined feature used in overall prediction by SVM.

A variety of studies have identified orlistat as a drug against obesity[88–90], acarbose as a drug against diabetes[91,92], and afatinib as a drug against diverse cancer[93,94]. Docking interactions of all these drugs have been shown in Figs. 6 and 7, and 8 respectively.

Molecular dynamic simulation

After having top-docked hits against Src kinase, we further investigated the structural changes in best complexes for a time span of 300 ns. We considered key parameters like root mean square deviation (RMSD), root mean square fluctuation (RMSF), radius of gyration (RoG) and hydrogen bonding between proposed drugs and receptor Src kinase to check possible structural alterations and dynamic behavior in simulation time. Along with the top drug hits, we also considered already FDA-approved inhibitor of tyrosine kinase protein Src kinase, dasatinib, to compare its binding affinity with the proposed drugs. It was taken as a control and showed comparably less stability and compactness than the identified drugs of this study.

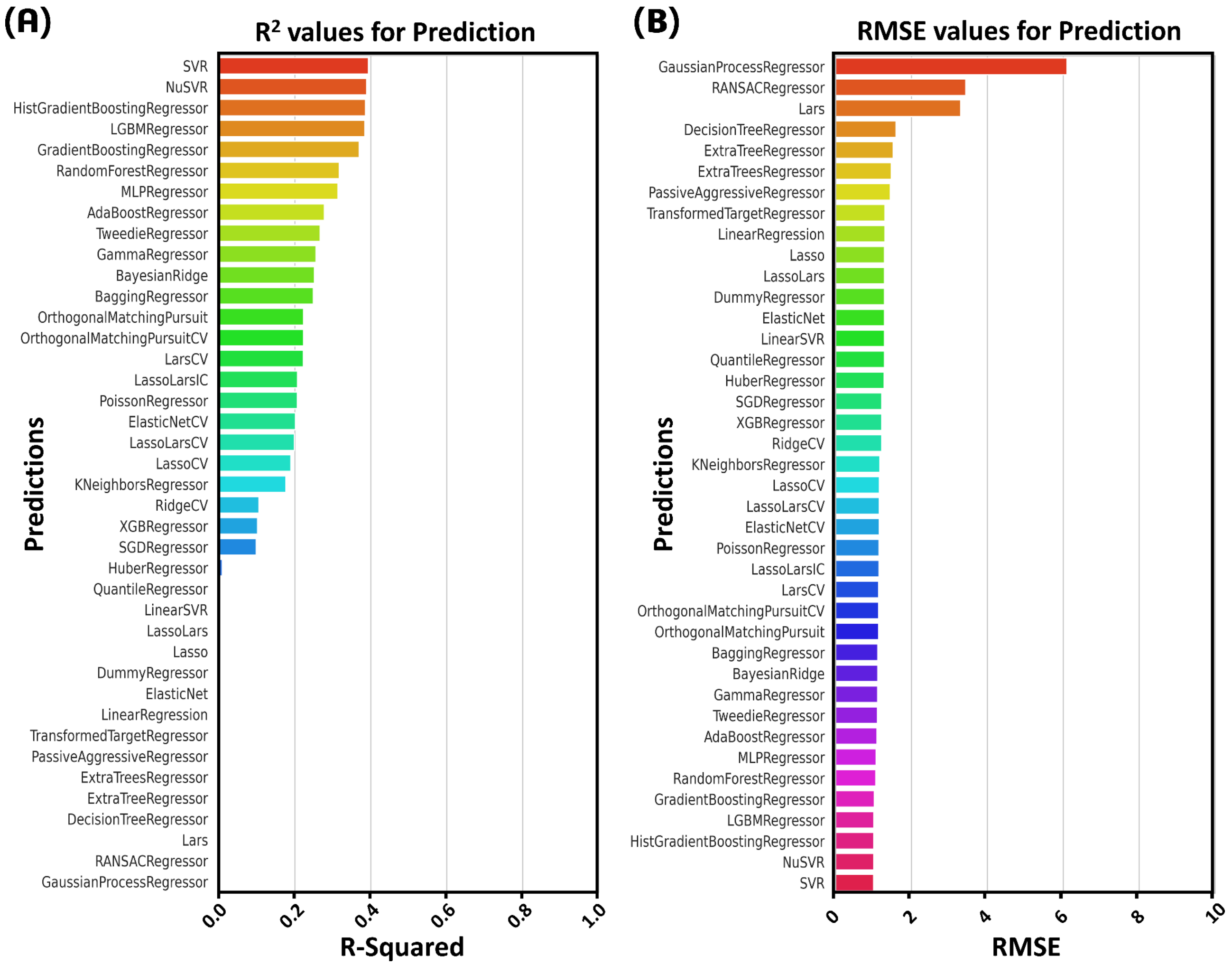

Fig. 5. **(A)** A bar graph representing the performance validation of various ML models based on their R-squared values **(B)** A bar graph representing the performance validation of various ML models based on their RMSE values.

Compound name	Pubchem ID	S-Score	RMSD
Orlistat	3,034,010	− 8.71	1.67
Acarbose	41,774	− 7.92	1.95
Afatinib	57,519,523	− 7.72	1.60
Control (Dasatinib)	3,062,316	− 6.86	2.52

Table 1. Represents the top 3 drugs out of 1040, along with the control, arranged according to the previously indicated criteria.

Root mean square deviation

To evaluate the stability and binding conformation of simulated complex, Root Mean Square Deviation (RMSD) is widely used technique categorized as a post-simulation trajectory analysis. In this regard, the overall duration of MD simulation is also important for analyzing receptor-drug interactions. In this investigation, we conducted a thorough 300 ns of simulation for the better estimation of drug stability while binding in receptor's active site[95]. The dasatinib- Src kinase complex, which was taken as control, showed comparably higher deviations than all of our proposed drugs. Its overall RMSD values deviated up to 6 Å, with some stability at 4 Å for first 80 ns. After that, it showed some irregularities in the peak for the rest of the simulation process (Fig. 9A). On the other hand, RMSD values for orlistat-Src kinase complex showed least deviation at around 2 Å for whole of the simulation time, indicating highest stability (Fig. 9B). Likewise, the acarbose- Src kinase complex also exhibited good stability for the whole 300 ns simulation as its RMSD values showed stable behavior at around 2.5 Å (Fig. 9C). Finally, the complex between afatinib and Src kinase also demonstrated low deviations as the RMSD values kept

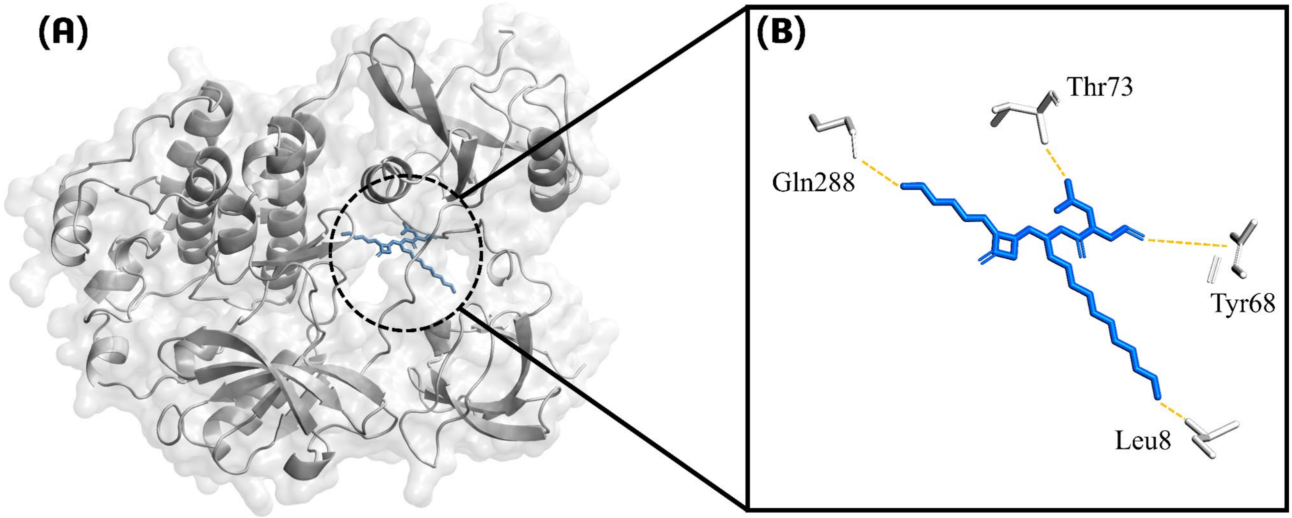

Fig. 6. Molecular interaction of the Src-orlistat complex **(A)** Overview of the Src enzyme (gray surface representation) with the bound orlistat molecule (blue stick representation) within the active site. The black box highlights the active site of the Src enzyme. (B) A zoomed-in view of the active site detailing the interactions between orlistat and the critical amino acid residues of Src. Orlistat is shown in blue stick representation, with interacting residues depicted in stick form (gray) and labeled accordingly.

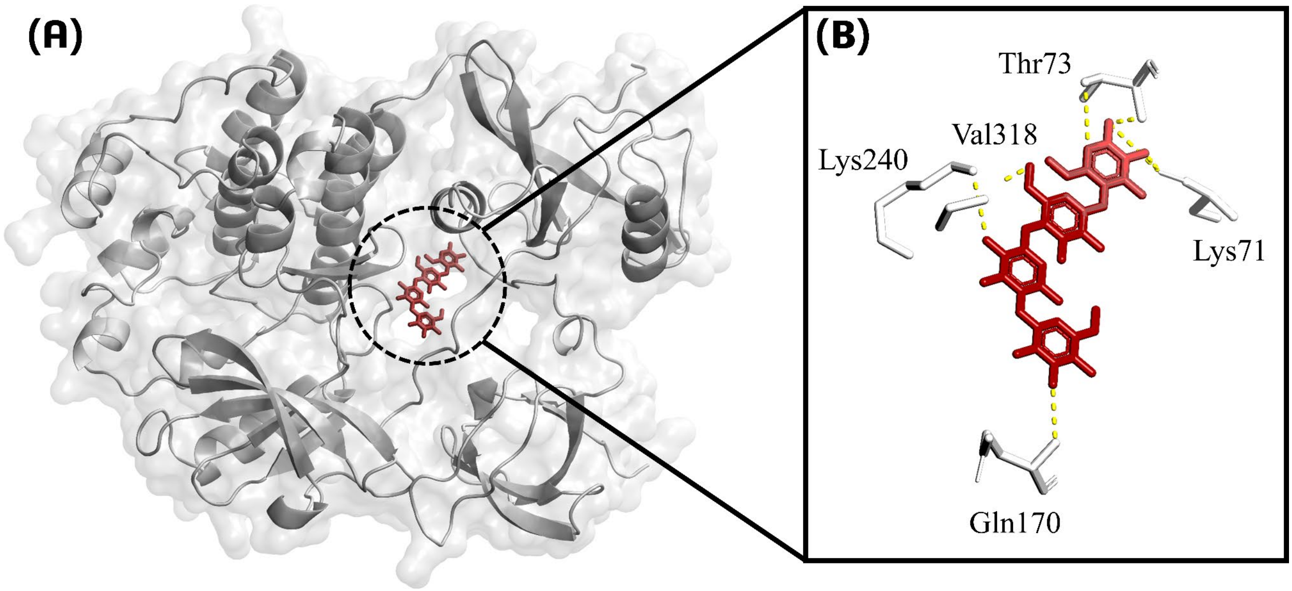

Fig. 7. Molecular interaction of the Src-acarbose complex **(A)** Overview of the Src enzyme (gray surface representation) with the bound acarbose molecule (red stick representation) within the active site. The black box highlights the active site of the Src enzyme. (B) A zoomed-in view of the active site detailing the interactions between acarbose and the critical amino acid residues of Src. Acarbose is shown in red stick representation, with interacting residues depicted in stick form (gray) and labeled accordingly.

below 3 Å, which indicates reliable stability than the control (Fig. 9D). The overall RMSD values indicated that all of the proposed drugs form significantly stable behavior with the receptor SRC KINASE, as compared to the dasatinib, indicating their potential against inflammatory diseases.

Root mean square fluctuation

As RMSD is the stability measure of the whole complex, the Root Mean Square Fluctuation (RMSF) represents the stability and flexibility of the simulated complex on residual level. It measures the flexibility of each residue according to it fluctuation from the mean, where high fluctuation indicates higher flexibility (i.e. lower stability) and vice versa[96]. In this investigation, all the complexes, including control, showed surprisingly very stable

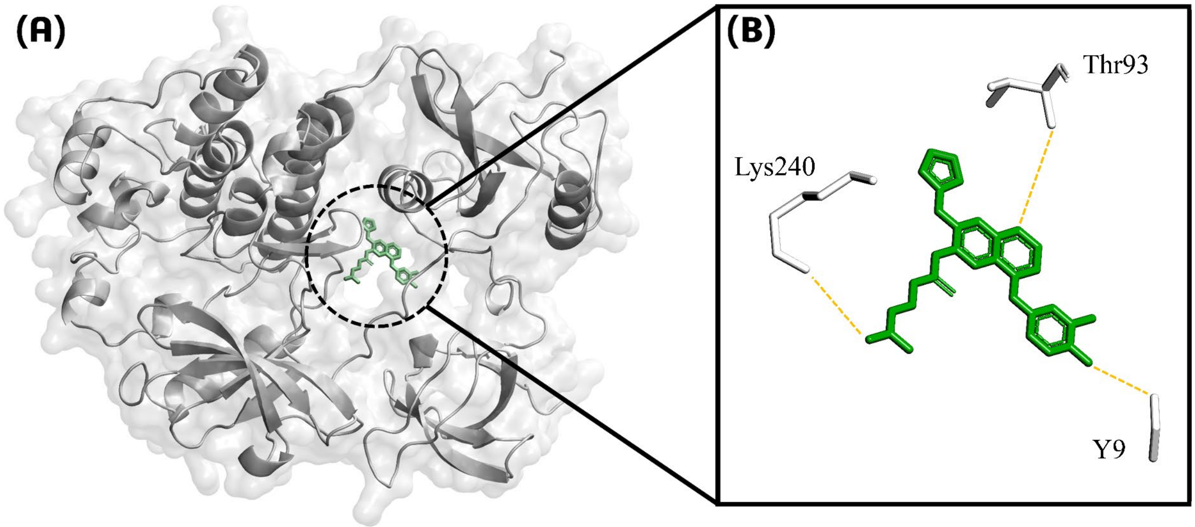

Fig. 8. Molecular interaction of the Src-afatinib complex **(A)** Overview of the Src enzyme (gray surface representation) with the bound afatinib molecule (green stick representation) within the active site. The black box highlights the active site of the Src enzyme. (B) A zoomed-in view of the active site detailing the interactions between afatinib and the critical amino acid residues of Src. Afatinib is shown in green stick representation, with interacting residues depicted in stick form (gray) and labeled accordingly.

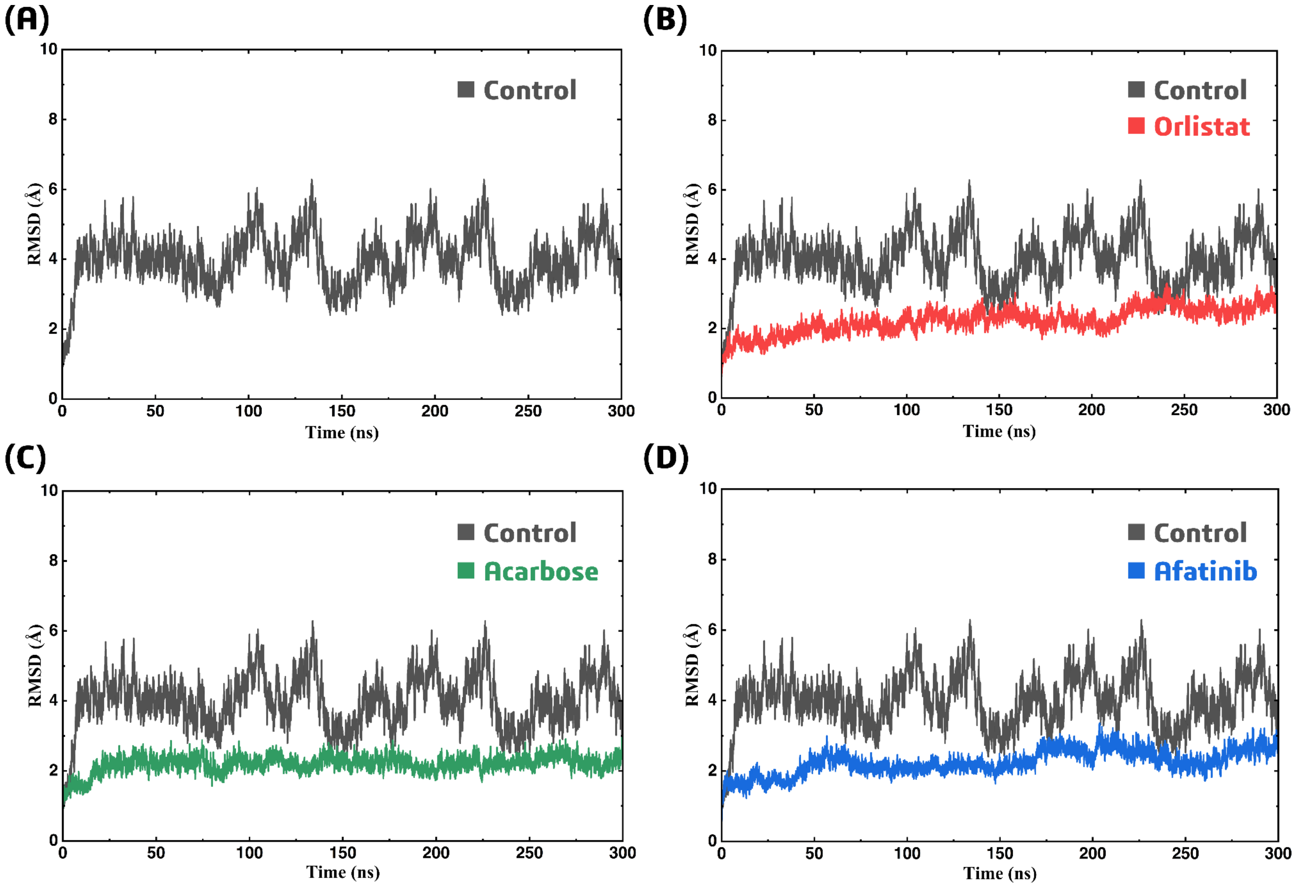

Fig. 9. **(A)** Representing Root Mean Square Deviation of control **(B)** Representing Root Mean Square Deviation of orlistat-Src kinase complex **(C)** Representing Root Mean Square Deviation of acarbose-Src kinase complex **(D)** Representing Root Mean Square Deviation of afatinib-Src kinase complex.

behavior throughout the whole simulation (Fig. 10). All of the complexes exhibited least fluctuations and kept stable at around 1.7 Å. In all the complexes, first 5 residues showed minimal fluctuations, at about 8 Å for control and 5 Å for each proposed drug, indicating the start of simulation process. Likewise, the last 5 residues of all the complexes also exhibited slight fluctuations around 4 Å approximately, indicating the completion of simulation. The overall RMSF results suggested that all the complexes showed stable behavior throughout the simulation, which must be validated by further analysed like RoG and hydrogen bond analysis.

Radius of gyration

After evaluating stability using RMSD and RMSF analyses, we, then, validate the efficacy of the proposed drugs by looking into their compactness and conformational stability, using a key metric names radius of gyration (RoG)[97]. To conclude the RoG analysis results, we delve deeper into the resulted values of this analysis. The dasatinib-Src complex (control) exhibited comparably less conformational stability than all of the proposed drugs and the RoG values were also slightly higher at around 40.3 Å (Fig. 11A). Conversely, the orlistat-Src kinase complex demonstrated very stable RoG values at around 38.9, indicating the highest compactness of the complex throughout 300 ns simulation (Fig. 11B). Likewise, highly stable RoG values of acarbose-Src kinase complex between 38.8 Å to 38.9 Å indicated its highest compactness throughout simulation (Fig. 11C). At the end, afatinib complexed with Src kinase also demonstrated highly stable behavior as its values showed less deviation at around 38.9 Å (Fig. 11D). All of these RoG values suggested that the proposed drugs complexed with the receptor Src kinase exhibited better conformational stability than that of control, which further validated their potential and efficacy to inhibit Src kinase for inflammatory diseases' treatment.

Hydrogen bond analysis

Hydrogens and hydrophobic contacts at the interface greatly improve the interaction of protein-ligand complexes[98]. We determined all of the hydrogen bonds produced in each system according to predetermined standards to assess these interactions at the atomic level. A hydrogen bond was deemed created if the hydrogen donor-acceptor angle was within 30 degrees and the donor-acceptor distance was within 0.35 nm[99]. A time-dependent study of hydrogen bonding is shown in Fig. 12, demonstrating that all three complexes had stronger hydrogen bonding networks than the control. This analysis fairly indicates that all three proposed complexes have a significant binding affinity, enabling precise binding to Src kinase, which is necessary for their potential

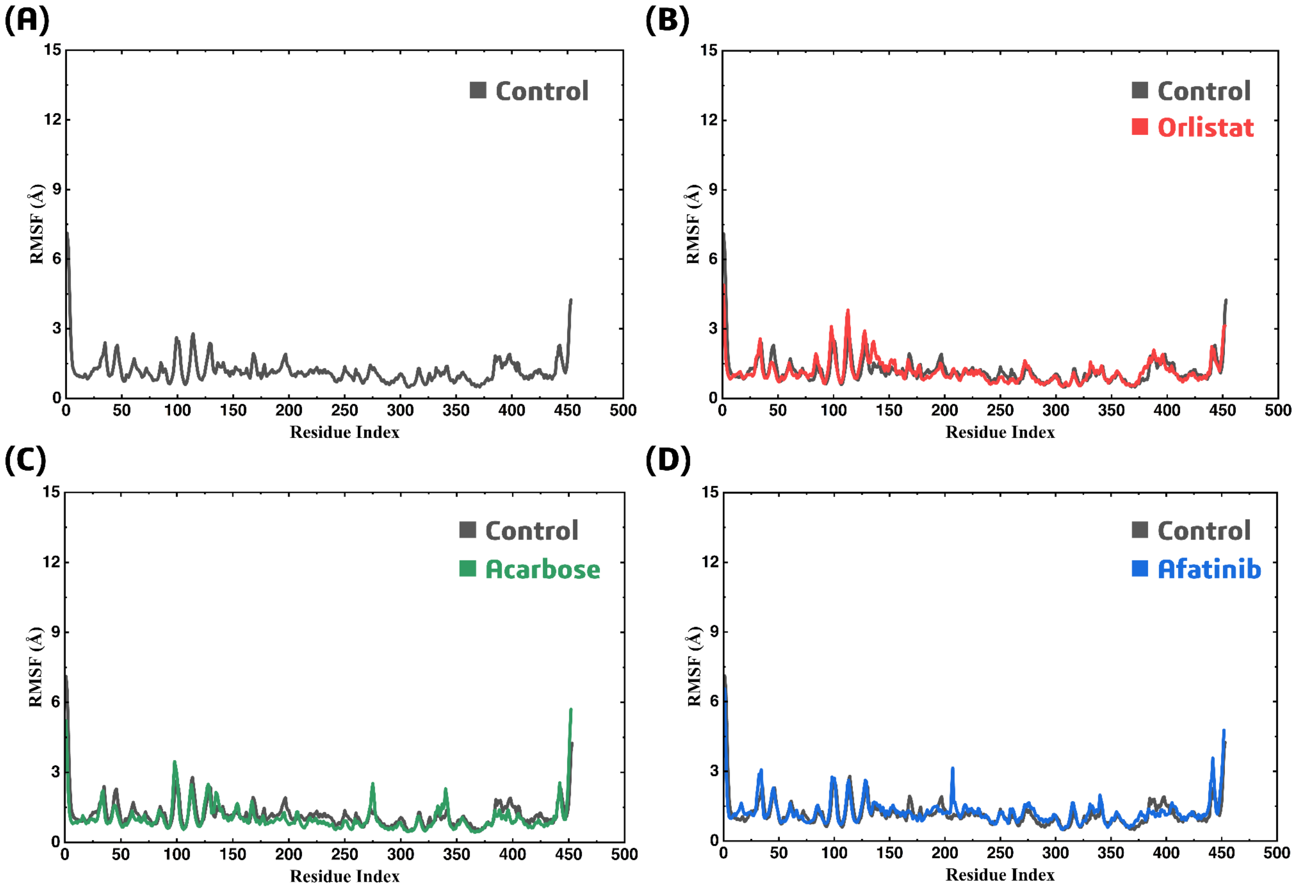

Fig. 10. **(A)** Representing Root Mean Square Fluctuation of control **(B)** Representing Root Mean Square Fluctuation of orlistat-Src kinase complex **(C)** Representing Root Mean Square Fluctuation of acarbose-Src kinase complex **(D)** Representing Root Mean Square Fluctuation of afatinib-Src kinase complex.

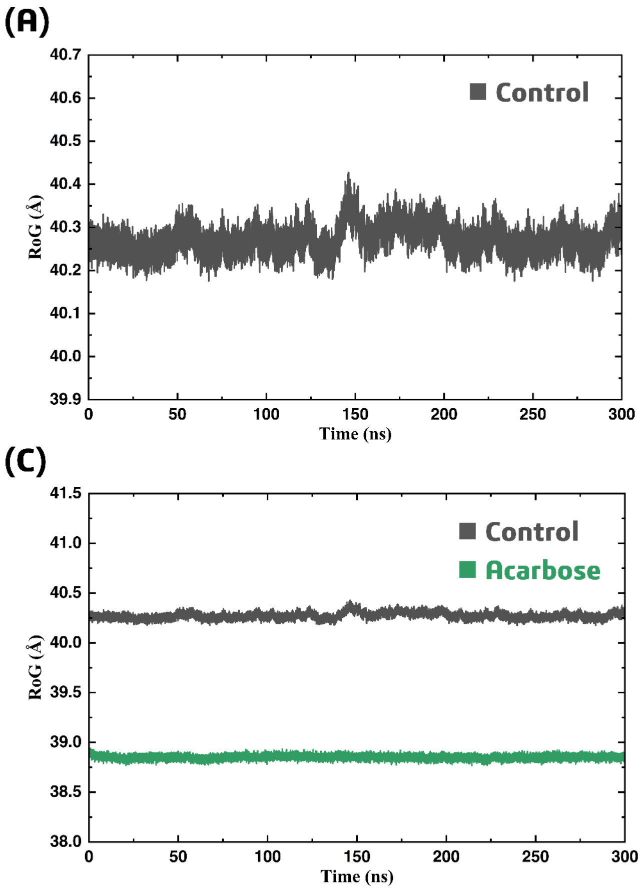

Fig. 11. **(A)** Representing Radius of Gyration of control **(B)** Representing Radius of Gyration of orlistat-Src kinase complex **(C)** Representing Radius of Gyration of acarbose-Src kinase complex **(D)** Representing Radius of Gyration of afatinib-Src kinase complex.

use against inflammatory disease. The presence of hydrogen bonds consistently throughout the simulation indicates strong and stable interactions. Specifically, the analysis revealed that each complex maintained a considerable number of hydrogen bonds, underscoring their substantial binding affinity. This strong hydrogen bonding network is crucial for maintaining the secondary structure of the protein-ligand complexes, thereby enhancing their stability and therapeutic potential.

Free energy calculations

The MM-GBSA and MM-PBSA strategies are commonly employed to re-score the ligand's docked conformation, identifying the most accurate ligand orientation and the optimal binding free energy. The estimated overall MMGBSA and MMPBSA values for control were – 13.7785 ± 5.8058 and – 2.3095 ± 5.7438. For the orlistat, acarbose and afatinib complexes, the MMGBSA values were calculated to be – 33.4743 ± 3.8908, – 19.5455 ± 5.4702, and – 36.4944 ± 5.4929 respectively (Table 2), and MMPBSA values were – 13.5671 ± 4.2658, – 12.0854 ± 4.9705, -and 16.9450 ± 4.0926, respectively (Table 3). These results demonstrated that proposed drugs complexes with the receptor Src kinase showed lower binding free energy than control, suggesting highest binding affinity and stability between them.

Van der Waals energy (vdW), electrostatic energy (EEL), surface energy (ESURF), Poisson-Boltzmann energy (EPB), polar energy (EnPolar), Generalized Born energy (EGB), and surface energy (ESURF) are among the components of binding free energy. These elements work together to shed light on the interactions between molecules. Overall, the MM-GBSA and MM-PBSA studies ranked the three potential inhibitors, including orlistat, acarbose, and afatinib, higher in binding affinity than the control (Fig. 13), confirming their strong binding against Src kinase, which is linked with various inflammatory diseases.

Analysis of toxicity profile

Finally, we calculated the toxicity profile of the proposed drugs, including orlistat, acarbose, afatinib and control (dasatinib), which demonstrated varying degree of toxicity as depicted in Fig. 14. Along with the control, afatinib also showed surprising toxicity concerns as its NR-AhR marker peaked at 0.75, demonstrating its binding and activation of aryl hydrocarbon receptor (AhR) that suggests it could behave as xenobiotic or environmental toxicant. Likewise, its marker's peak with SR-ARE at about 0.75 indicated it's binding with nuclear factor erythroid 2-related factor 2, which may lead to oxidative stress. Furthermore, its exceeded threshold with SR-

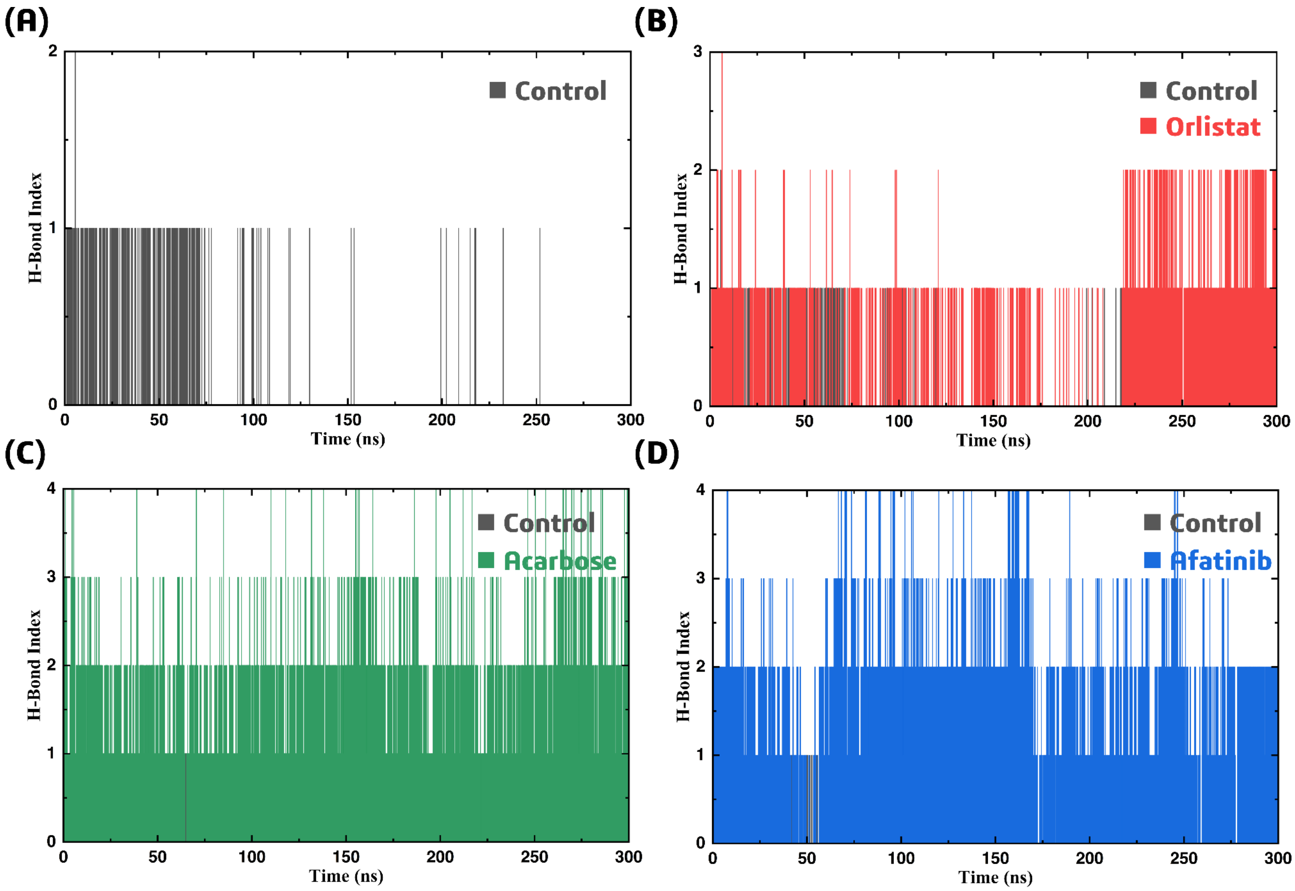

Fig. 12. **(A)** Representing hydrogen bonding of control **(B)** Representing hydrogen bonding of orlistat-Src kinase complex **(C)** Representing hydrogen bonding of acarbose-Src kinase complex **(D)** Representing hydrogen bonding of afatinib-Src kinase complex.

Drug name	vdW	EEL	EGB	ESURF	TOTAL + STD (Kcal/mol)
Orlistat	− 47.3805	− 5.0151	24.9509	− 6.0295	− 33.4743 ± 3.8908
Acarbose	− 36.2241	− 12.0929	34.2599	− 5.4884	− 19.5455 ± 5.4702
Afatinib	− 49.1236	− 128.6558	146.9905	− 5.7056	− 36.4944 ± 5.4929
Control	− 28.8002	− 2.2707	22.6055	− 3.3131	− 13.7785 ± 5.8058

Table 2. Represents the binding free energies of each complex, calculated by MMGBSA.

Drug name	vdW	EEL	EPB	EnPolar	TOTAL + STD (Kcal/mol)
Orlistat	− 47.3805	− 5.0151	27.3354	− 34.3541	− 13.5671 ± 4.2658
Acarbose	− 36.2241	− 12.0929	34.8462	− 28.2797	− 12.0854 ± 4.9705
Afatinib	− 49.1236	− 128.6558	149.5679	− 31.6261	− 16.9450 ± 4.0926
Control	− 28.8002	− 2.2707	12.7330	− 19.8464	− 2.3095 ± 5.7438

Table 3. Represents binding free energies of each complex, calculated by MMPBSA.

MMP (0.69) and SR-p53 (0.58) demonstrated the mitochondrial toxicity, DNA damage and other genotoxic effects. In contrast, orlistat and acarbose displayed more balanced toxicity profiles, with no significant peaks exceeding 0.5 threshold, suggesting these two drugs as better therapeutic options. This analysis provides critical insights for guiding future clinical development and potential therapeutic applications of these compounds.

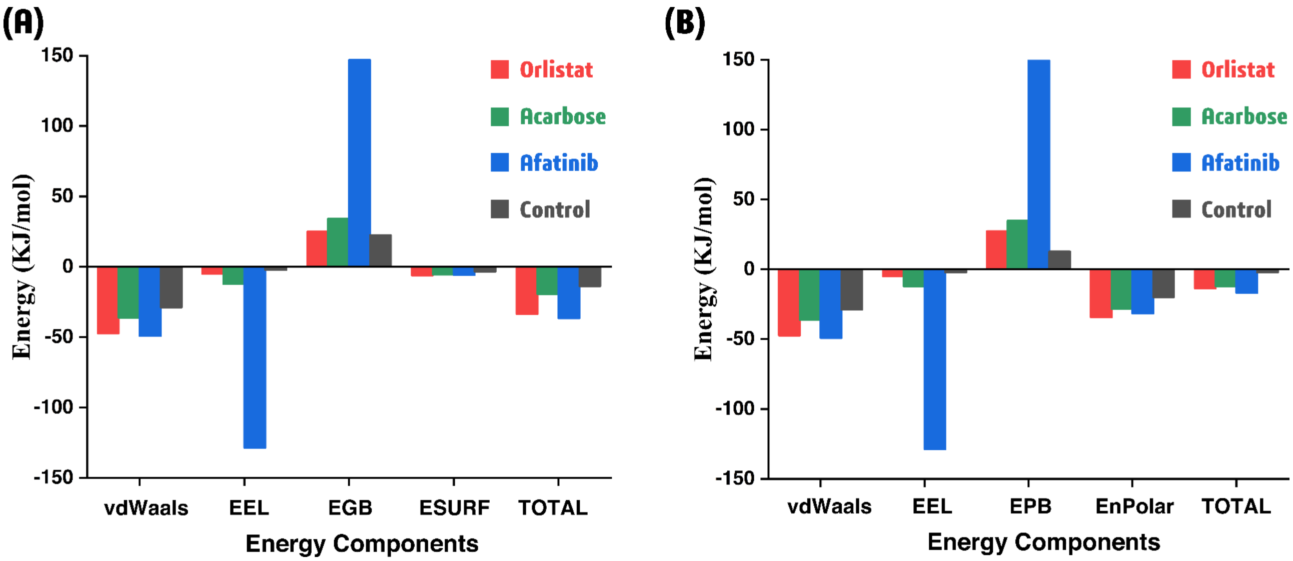

Fig. 13. **(A)** A bar graph representing average energy of relevant components, estimated by MMGBSA **(B)** A bar graph representing average energy of relevant components, estimated by MMPBSA.

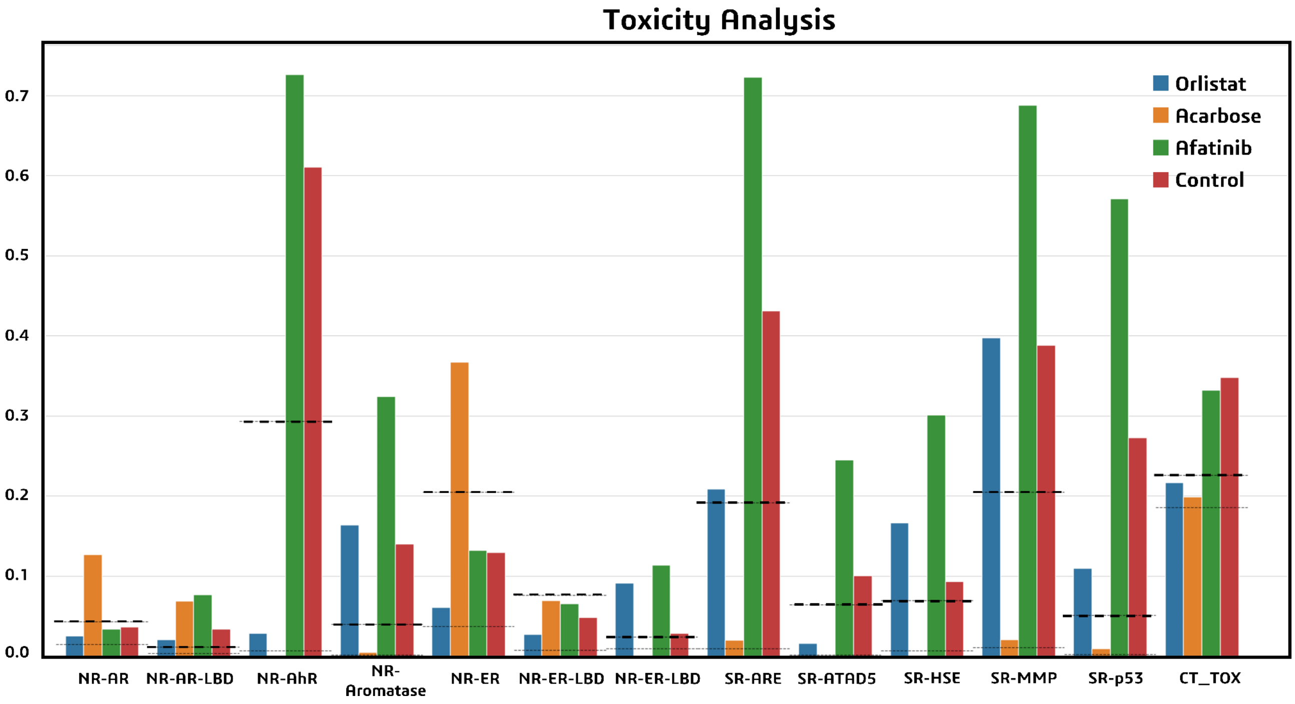

Fig. 14. A barplot representing the toxicity profile of all the three proposed drugs including orlistat, acarbose and afatinib, along with the control (dasatinib).

Discussion

The primary aim of our study was to propose a comprehensive and efficient approach by combining both structure-based drug designing and the use of machine learning in the field of drug discovery. This integrated approach was utilized to target a crucial non-receptor protein kinase, Src kinase, reported to have its role in causing and progressing various inflammatory diseases including systemic lupus erythematosus (SLE)[3], rheumatoid arthritis[2], atherosclerosis[4], psoriasis[6], pemphigoid diseases[8], osteoporosis[7], and inflammatory bowel disease (IBD)[5]. Because of its involvement in inflammatory signaling pathways, Src kinase is a crucial therapeutic target. However, the current FDA-approved inhibitors, such as dasatinib, have drawbacks such as poor selectivity, specificity, and off-target effects[30]. As studies reported that both mutations and overexpression of Src kinase have been involved in inflammatory diseases, we aimed to inhibit its activity by identifying potential inhibitors. The pipeline began with the retrieval of 3,570 Src targeting bioactive compounds from ChEMBL database, which underwent a variety of refinement phases including removal of duplicates, ADMET and PAINS analyses, and substructures elimination. This comprehensive refinement resulted in the selection of 586 compounds from the whole bioactive dataset for further processing. This thorough refinement emphasized that the shortlisted compounds were highly accurate, which were further processed to train different machine learning models including random forest, decision tree, support vector machine, and k-nearest neighbors. For this purpose, these 586 compounds were converted into machine-recognized binary format so that each model can recognize and read the respective features. After having binarized dataset, feature selection techniques, such as permutation importance analysis and recursive feature elimination (RFE), further refined it to include only 92 highly correlated features, enhancing model interpretability and performance. Following this, all of the four machine learning models were trained and tested in 80/20 fashion, based on the highly refined and accurate bioactive dataset. Among the models tested, Support vector machine (SVM) model stood out as the best performing machine learning model as its R^2 score (0.38) and RMSE (0.1) were significantly reliable as compared to random forest (R^2: 0.30, RMSE: 0.11), K-NN (R^2: 0.25, RMSE: 0.13), and decision tree (R^2: − 0.35, RMSE: 0.18). Subsequently, all of the four tested models were further evaluated for their performance validation by comparing them with the other 37 machine learning models. The results indicated that SVM outperformed all of the compared models, suggesting its high accuracy. Following this, the screening of already available FDA-approved library of 1040 drug using trained SVM model resulted in identifying the top 51 compounds, having pIC_{50} values over 6.5, which is usually considered in a range of reliable drug potency. These shortlisted compounds were further screened using molecular docking, which provided top 3 compounds, orlistat (S-Score: − 8.71, RMSD: 1.67), acarbose (S-Score: − 7.92, RMSD: 1.95), and afatinib (S-Score: − 7.72, RMSD: 1.60), as output due to their lower S-scores and RMSD values. Compared to an FDA-approved Src inhibitor dasatinib (S-Score: − 6.86, RMSD: 2.52), all of these compounds demonstrated superior binding efficiency, underscoring their potential as Src inhibitors. These results represents that the chosen compounds have strong binding affinity and potential to inhibit Src kinase. To validate the docking results, all-atom molecular dynamic simulation of 300 ns was conducted, which assessed RMSD, RMSF, RoG and H-Bond analysis to demonstrate results. The comparably lower deviations in the RMSD, RMSF, and RoG values of all the potent leads demonstrated their higher binding potential with Src. Furthermore, MMGBSA and MMPBSA evaluation concluded that orlistat (− 33.47), acarbose (− 19.55), and afatinib (− 36.49) showed comparatively better stability and binding affinity than dasatinib (− 13.78). However, toxicity analysis eliminated afatinib as it exhibited potential safety issue related to its interactions with NR-AhR (0.75), SR-ARE (0.75), SR-MMP (0.69) and SR-p53 (0.58), indicating its various toxicity effects including environmental and mitochondrial toxicity. Orlistat and acarbose stood out as the safer and non-toxic inhibitors of Src kinase protein. One of the key strengths of our investigation is the uniqueness of its methodology as we employed multiple computational techniques to enhance the reliability of this study. The utilized machine learning algorithms, including RF, decision tree, K-NN and SVM, have been widely recognized for their accuracy, efficiency and efficacy[100,101]. The translational potential of our findings is further enhanced by the repurposing of FDA-approved drugs, as these compounds already have established safety profiles and pharmacokinetics. Despite these strengths, this research still have some limitations as possible inconsistencies in the retrieved bioactive data and intrinsic constraints of molecular simulation studies, which may not completely depict the complexity of in-vivo experiments. We suggest the further in-vivo and in-vitro validation of our proposed drugs before formally use in clinical applications. Our results demonstrate the potential of acarbose and orlistat as Src kinase inhibitors with future applications in the management of inflammatory diseases. Moreover, the employed strategy can also be applicable for other drug targets as well, which broads the scope of our employed strategy. This approach also highlights the use of machine learning applications and drug repurposing for the identification of novel Src kinase inhibitors to prevent the progression of inflammatory diseases.

Conclusion

As genetic mutations and overexpression of tyrosine protein kinase Src is involved in the progression of various inflammatory diseases, we employed a combined machine learning and structure-based drug repurposing strategy to inhibit Src kinase activity with the aim to prevent linked inflammatory diseases. By identifying orlistat and acarbose as promising candidates using an integrated approach, we addressed crucial off-target binding, toxicity and selectivity problems of previously identified inhibitors. The computational strategies used in this study including machine learning, molecular docking and molecular simulation, revealed accurate and strong binding affinities and stability of the proposed compounds, highlight their potential as effective Src inhibitors. This study also suggested the potential use of machine learning-based computational approaches in the field of drug discovery and proposed orlistat and acarbose as potent Src kinase inhibitors, which require further experimental approval.

Data availability

All data generated or analyzed during this study are included in this published article.

Received: 3 September 2024; Accepted: 17 December 2024
Published online: 13 January 2025

References

1. David, T., Ling, S. & Barton, A. Genetics of immune-mediated inflammatory diseases. *Clin. Exp. Immunol.* **193**(1), 3–12 (2018).

2. Siebert, S. et al. Cytokines as therapeutic targets in rheumatoid arthritis and other inflammatory diseases. *Pharmacol. Rev.* **67**(2), 280–309 (2015).
3. Aringer, M. Inflammatory markers in systemic lupus erythematosus. *J. Autoimmun.* **110**, 102374 (2020).
4. Ross, R. Atherosclerosis—an inflammatory disease. *N. Engl. J. Med.* **340**(2), 115–126 (1999).
5. Baumgart, D. C. & Carding, S. R. Inflammatory bowel disease: cause and immunobiology. *Lancet* **369**(9573), 1627–1640 (2007).
6. Davidovici, B. B. et al. Psoriasis and systemic inflammatory diseases: potential mechanistic links between skin disease and co-morbid conditions. *J. Investig. Dermatol.* **130**(7), 1785–1796 (2010).
7. Lacativa, P. G. S. & Farias, M. L. F. Osteoporosis and inflammation. *Arqu. Brasil Endocrinol. Metabol.* **54**, 123–132 (2010).
8. Schmidt, E. & Zillikens, D. Pemphigoid diseases. *Lancet* **381**(9863), 320–332 (2013).
9. Gabriel, S. E. The epidemiology of rheumatoid arthritis. *Rheum. Dis. Clin. N Am.* **27**(2), 269–281 (2001).
10. Ramos-Casals, M. et al. *Sjögren's Syndrome. A Clinician's Pearls and Myths in Rheumatology* 107–130 (Springer, 2009).
11. Taylor, T. et al. Causes of death among individuals with systemic lupus erythematosus by race and ethnicity: a population-based study. *Arthritis Care Res.* **75**(1), 61–68 (2023).
12. Roifman, I. et al. Chronic inflammatory diseases and cardiovascular risk: a systematic review. *Can. J. Cardiol.* **27**(2), 174–182 (2011).
13. Dave, M. et al. Opportunistic infections due to inflammatory bowel disease therapy. *Inflamm. Bowel Dis.* **20**1), 196–212 (2014).
14. Parke, A. & Parke, D. V. The pathogenesis of inflammatory disease: surgical shock and multiple system organ failure. *InflammoPharmacology* **3**(2), 149–168 (1995).
15. Cassell, G. H. Infectious causes of chronic inflammatory diseases and cancer. *Emerg. Infect. Dis.* **4**(3), 475–487 (1998).
16. Packey, C. D. & Sartor, R. B. Interplay of commensal and pathogenic bacteria, genetic mutations, and immunoregulatory defects in the pathogenesis of inflammatory bowel diseases. *J. Intern. Med.* **263**(6), 597–606 (2008).
17. Singh, N. N. & Ramji, D. P. Protein kinase CK2, an important regulator of the inflammatory response? *J. Mol. Med.* **86**(8), 887–897 (2008).
18. Roskoski, R. Src protein-tyrosine kinase structure, mechanism, and small molecule inhibitors. *Pharmacol. Res.* **94**, 9–25 (2015).
19. Boggon, T. J. & Eck, M. J. Structure and regulation of src family kinases. *Oncogene* **23**(48), 7918–7927 (2004).
20. Turro, E. et al. A dominant gain-of-function mutation in universal tyrosine kinase SRC causes thrombocytopenia, myelofibrosis, bleeding, and bone pathologies. *Sci. Transl. Med.* **8**(328), 328ra30 (2016).
21. Yu, C. C. K. et al. Lupus-like kidney disease in mice deficient in the src family tyrosine kinases Lyn and Fyn. *Curr. Biol.* **11**(1), 34–38 (2001).
22. Shao, W. H. & Cohen, P. L. The role of tyrosine kinases in systemic lupus erythematosus and their potential as therapeutic targets. *Expert Rev. Clin. Immunol.* **10**(5), 573–582 (2014).
23. Reddy, M. A. et al. Role of src tyrosine kinase in the atherogenic effects of the 12/15-lipoxygenase pathway in vascular smooth muscle cells. *Arterioscler. Thromb. Vasc. Biol.* **29**(3), 387–393 (2009).
24. Yang, L. & Yan, Y. Protein kinases are potential targets to treat inflammatory bowel disease. *World J. Gastrointest. Pharmacol. Ther.* **5**(4), 209–217 (2014).
25. Szilveszter, K. P., Németh, T. & Mócsai, A. Tyrosine kinases in autoimmune and inflammatory skin diseases. *Front. Immunol.* **10**, 1862 (2019).
26. Sohraby, F. et al. A boosted unbiased molecular dynamics method for predicting ligands binding mechanisms: probing the binding pathway of dasatinib to src-kinase. *Bioinformatics* **36**(18), 4714–4720 (2020).
27. Khamouli, S. et al. Comprehensive in silico discovery of c-Src tyrosine kinase inhibitors in cancer treatment: a unified approach combining pharmacophore modeling, 3D QSAR, DFT, and molecular dynamics simulation. *J. King Saud Univ. Sci.* **36**(3), 103076 (2024).
28. Rivera-Torres, J., San, E. & José Src tyrosine kinase inhibitors: new perspectives on their immune, antiviral, and senotherapeutic potential. *Front. Pharmacol.* **10**, 1011 (2019).
29. Sato, H. et al. SRC family kinase inhibition targets YES1 and YAP1 as primary drivers of lung cancer and as mediators of acquired resistance to ALK and epidermal growth factor receptor inhibitors. *JCO Precis Oncol.* **6**, e2200088 (2022).
30. Puls, L. N., Eadens, M. & Messersmith, W. Current status of SRC inhibitors in solid tumor malignancies. *Oncologist* **16**(5), 566–578 (2011).
31. Scholz, C. et al. DOCKTITE a highly versatile step-by-step workflow for covalent docking and virtual screening in the molecular operating environment. *J. Chem. Inf. Model.* **55**(2), 398–406 (2015).
32. Mabrouk, M. S. Discovering best candidates for Hepatocellular Carcinoma (HCC) by in-silico techniques and tools. *Int. J. Bioinform Res. Appl.* **8**(1–2), 141–152 (2012).
33. Gaulton, A. et al. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **40**(D1), D1100–D1107 (2011).
34. Bento, A. P. et al. An open source chemical structure curation pipeline using RDKit. *J. Cheminform.* **12**(1), 51 (2020).
35. Proctor, W. R. et al. Utility of spherical human liver microtissues for prediction of clinical drug-induced liver injury. *Arch. Toxicol.* **91**(8), 2849–2863 (2017).
36. Siddique, F. et al. Revisiting methotrexate and phototrexate Zinc15 library-based derivatives using deep learning in-silico drug design approach. *Front. Chem.* **12**, 1380266 (2024).
37. Wang, J. & Skolnik, S. Recent advances in physicochemical and ADMET profiling in drug discovery. *Chem. Biodivers.* **6**(11), 1887–1899 (2009).
38. Bisong, E. *Building Machine Learning and deep Learning Models on Google Cloud Platform* (Apress, 2019).
39. Baell, J. B. & Holloway, G. A. New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. *J. Med. Chem.* **53**(7), 2719–2740 (2010).
40. Sydow, D. et al. TeachOpenCADD: a teaching platform for computer-aided drug design using open source packages and data. *J. Cheminform.* **11**(1), 29 (2019).
41. Brenk, R. et al. Lessons learnt from assembling screening libraries for drug discovery for neglected diseases. *ChemMedChem* **3**(3), 435–444 (2008).
42. Yap, C. W. PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. *J. Comput. Chem.* **32**(7), 1466–1474 (2011).
43. Snider, L. A. & Swedo, S. E. PANDAS: current status and directions for research. *Mol. Psychiatry*. **9**(10), 900–907 (2004).
44. Nayarisseri, A. et al. Artificial intelligence, big data and machine learning approaches in precision medicine & drug discovery. *Curr. Drug Targets* **22**(6), 631–655 (2021).
45. Cano, G. et al. Automatic selection of molecular descriptors using random forest: application to drug discovery. *Expert Syst. Appl.* **72**, 151–159 (2017).
46. Arian, R. et al. Protein kinase inhibitors' classification using K-Nearest neighbor algorithm. *Comput. Biol. Chem.* **86**, 107269 (2020).
47. Maltarollo, V. G. et al. Advances with support vector machines for novel drug discovery. *Expert Opin. Drug Discov*. **14**(1), 23–33 (2019).
48. Bos, P. H. et al. AutoDesigner, a De Novo Design Algorithm for rapidly exploring large chemical space for lead optimization: application to the design and synthesis of d-Amino acid oxidase inhibitors. *J. Chem. Inf. Model.* **62**(8), 1905–1915 (2022).
49. Jaskowiak, P. A. et al. A comparative study on the use of correlation coefficients for redundant feature elimination. In *2010 Eleventh Brazilian Symposium on Neural Networks* (IEEE, 2010).

50. Raza, A. et al. AIPs-SnTCN: Predicting anti-inflammatory peptides using fasttext and transformer encoder-based hybrid word embedding with self-normalized temporal convolutional networks. *J. Chem. Inf. Model.* **63**(21), 6537–6554 (2023).
51. Altmann, A. et al. Permutation importance: a corrected feature importance measure. *Bioinformatics* **26**(10), 1340–1347 (2010).
52. Muhamad Iqbal Januadi, P. & Vincent, A. Comparison of machine learning land use-land cover supervised classifiers performance on satellite imagery sentinel 2 using lazy predict library. *Indonesian J. Data Sci.* **4**(3), 183–189 (2023).
53. Lappalainen, K., Piliougine, M. & Spagnuolo, G. Experimental comparison between various fitting approaches based on RMSE minimization for photovoltaic module parametric identification. *Energy Convers. Manage.* **258**, 115526 (2022).
54. Colin Cameron, A. & Windmeijer, F. A. G. An R-squared measure of goodness of fit for some common nonlinear regression models. *J. Econometr.* **77**(2), 329–342 (1997).
55. Morris, G. M. & Lim-Wilby, M. Molecular docking. *Methods Mol. Biol.* **443**, 365–382 (2008).
56. Burley, S. K. et al. Protein Data Bank (PDB): the single global macromolecular structure archive. *Methods Mol. Biol.* **1607**, 627–641 (2017).
57. Gutti, G. et al. In-silico guided design, screening, and molecular dynamic simulation studies for the identification of potential SARS-CoV-2 main protease inhibitors for the targeted treatment of COVID-19. *J. Biomol. Struct. Dyn.* **42**(4), 1733–1750 (2024).
58. Liguori, N. et al. Molecular dynamics simulations in photosynthesis. *Photosynth Res.* **144**(2), 273–295 (2020).
59. Ten Brink, T. & Exner, T. E. pK a based protonation states and microspecies for protein–ligand docking. *J. Comput. Aided Mol. Des.* **24**, 935–942 (2010).
60. Ruiz-Carmona, S. et al. rDock: a fast, versatile and open source program for docking ligands to proteins and nucleic acids. *PLoS Comput. Biol.* **10**(4), e1003571 (2014).
61. Cuzzolin, A. et al. DockBench: an integrated informatic platform bridging the gap between the robust validation of docking protocols and virtual screening simulations. *Molecules* **20**(6), 9977–9993 (2015).
62. Castro-Alvarez, A., Costa, A. M. & Vilarrasa, J. The performance of several docking programs at reproducing protein-macrolide-like crystal structures. *Molecules* **22**, 1 (2017).
63. Butt, S. S. et al. Molecular docking using chimera and autodock vina software for nonbioinformaticians. *JMIR Bioinform. Biotechnol.* **1**(1), e14232 (2020).
64. Seeliger, D. & de Groot, B. L. Ligand docking and binding site analysis with PyMOL and Autodock/Vina. *J. Comput. Aided Mol. Des.* **24**(5), 417–422 (2010).
65. Singh, A. et al. Application of molecular dynamic simulation to study food proteins: a review. *Crit. Rev. Food Sci. Nutr.* **58**(16), 2779–2789 (2018).
66. Love, O. et al. Evaluating the accuracy of the AMBER protein force fields in modeling dihydrofolate reductase structures: misbalance in the conformational arrangements of the flexible loop domains. *J. Biomol. Struct. Dyn.* **41**(13), 5946–5960 (2023).
67. Junmei, W. Antechamber, an accessory software packagefor molecular mechanical calculations. *J. Chem. Inf. Comput. Sci.* (2001).
68. Loschwitz, J. et al. Dataset of AMBER force field parameters of drugs, natural products and steroids for simulations using GROMACS. *Data Brief.* **35**, 106948 (2021).
69. Kini, R. M. & Evans, H. J. Molecular modeling of proteins: a strategy for energy minimization by molecular mechanics in the AMBER force field. *J. Biomol. Struct. Dyn.* **9**(3), 475–488 (1991).
70. Liu, J., Li, D. & Liu, X. A simple and accurate algorithm for path integral molecular dynamics with the Langevin thermostat. *J. Chem. Phys.* **145**(2), 024103 (2016).
71. Lin, Y. et al. Application of Berendsen barostat in dissipative particle dynamics for nonequilibrium dynamic simulation. *J. Chem. Phys.* **146**(12), 124108 (2017).
72. Kräutler, V., van Gunsteren, W. F. & Hünenberger, P. H. A fast SHAKE algorithm to solve distance constraint equations for small molecules in molecular dynamics simulations. *J. Comput. Chem.* **22**(5), 501–508 (2001).
73. Möller, D. & Fischer, J. Vapour liquid equilibrium of a pure fluid from test particle method in combination with NpT molecular dynamics simulations. *Mol. Phys.* **69**(3), 463–473 (1990).
74. Harris, J. A. et al. GPU-Accelerated all-atom particle-mesh ewald continuous constant pH molecular dynamics in amber. *J. Chem. Theory Comput.* **18**(12), 7510–7527 (2022).
75. Roe, D. R. & Cheatham, T. E. PTRAJ and CPPTRAJ: software for processing and analysis of molecular dynamics trajectory data. *J. Chem. Theory Comput.* **9**(7), 3084–3095 (2013).
76. Rastelli, G. et al. Fast and accurate predictions of binding free energies using MM-PBSA and MM-GBSA. *J. Comput. Chem.* **31**(4), 797–810 (2010).
77. Kuhn, B. et al. Validation and use of the MM-PBSA approach for drug discovery. *J. Med. Chem.* **48**(12), 4040–4048 (2005).
78. Stoyanova, R. et al. Computational predictions of nonclinical pharmacokinetics at the drug design stage. *J. Chem. Inf. Model.* **63**(2), 442–458 (2023).
79. Wu, Z. et al. MoleculeNet: a benchmark for molecular machine learning. *Chem. Sci.* **9**(2), 513–530 (2018).
80. Sturm, N. et al. Application of bioactivity profile-based fingerprints for building machine learning models. *J. Chem. Inf. Model.* **59**(3), 962–972 (2019).
81. Cramer, R. D. et al. Virtual screening for R-groups, including predicted pIC50 contributions, within large structural databases, using Topomer CoMFA. *J. Chem. Inf. Model.* **48**(11), 2180–2195 (2008).
82. Hannaert, P. et al. Rat NKCC2/NKCC1 cotransporter selectivity for loop diuretic drugs. *Naunyn-Schmiedeberg Arch. Pharmacol.* **365**, 193–199 (2002).
83. Polanski, J., Bogocz, J. & Tkocz, A. The analysis of the market success of FDA approvals by probing top 100 bestselling drugs. *J. Comput. -Aided Mol. Des.* **30**, 381–389 (2016).
84. Masarweh, N. *Computational Modelling of Drugs for Alzheimer's Disease (Ad) and Applications on Artifical Neural Network Systems (Nets)* (University of Arkansas at Little Rock, 2021).
85. Hann, M. M. Molecular obesity, potency and other addictions in drug discovery. In *Multifaceted Roles of Crystallography in Modern Drug Discovery* 183–196 (Springer, 2015).
86. Ramírez, D. & Caballero, J. Is it reliable to take the molecular docking top scoring position as the best solution without considering available structural data?. *Molecules* **23**, 5 (2018).
87. Zhu, J. et al. Theoretical studies on the selectivity mechanisms of glycogen synthase kinase 3β (GSK3β) with pyrazine ATP-competitive inhibitors by 3D-QSAR, molecular docking, molecular dynamics simulation and free energy calculations. *Curr. Comput. -Aided Drug Des.* **16**(1), 17–30 (2020).
88. Henness, S. & Perry, C. M. Orlistat: a review of its use in the management of obesity. *Drugs* **66**(12), 1625–1656 (2006).
89. Hvizdos, K. M. & Markham, A. Orlistat: a review of its use in the management of obesity. *Drugs* **58**(4), 743–760 (1999).
90. Ballinger, A. & Peikin, S. R. Orlistat: its current status as an anti-obesity drug. *Eur. J. Pharmacol.* **440**(2), 109–117 (2002).
91. Balfour, J. A. & McTavish, D. Acarbose: an update of its pharmacology and therapeutic use in diabetes mellitus. *Drugs* **46**(6), 1025–1054 (1993).
92. Campbell, L. K., White, J. R. & Campbell, R. K. Acarbose: its role in the treatment of diabetes Mellitus. *Ann. Pharmacother.* **30**(11), 1255–1262 (1996).
93. Wind, S. et al. Clinical pharmacokinetics and pharmacodynamics of afatinib. *Clin. Pharmacokinet.* **56**(3), 235–250 (2017).
94. Vavalà, T. Role of afatinib in the treatment of advanced lung squamous cell carcinoma. *Clin. Pharmacol.* **9**, 147–157 (2017).
95. Salo-Ahen, O. M. et al. Molecular dynamics simulations in drug discovery and pharmaceutical development. *Processes* **9**(1), 71 (2020).

96. Maruyama, Y. et al. Analysis of protein folding simulation with moving root mean square deviation. *J. Chem. Inf. Model.* **63**(5), 1529–1541 (2023).
97. Liu, P. et al. Lubricant shear thinning behavior correlated with variation of radius of gyration via molecular dynamics simulations. *J. Chem. Phys.* **147**, 8 (2017).
98. Bizzarri, A. et al. Hydrogen bond analysis by MD simulation of copper plastocyanin at different hydration levels. *Chem. Phys.* **201**(2–3), 463–472 (1995).
99. Chen, C. et al. Hydrogen bonding analysis of glycerol aqueous solutions: a molecular dynamics simulation study. *J. Mol. Liquids* **146**(1–2), 23–28 (2009).
100. Hochreiter, S., Klambauer, G. & Rarey, M. Machine learning in drug discovery. *J. Chem. Inf. Model.* **58**(9), 1723–1724 (2018).
101. Priya, S. et al. Machine learning approaches and their applications in drug discovery and design. *Chem. Biol. Drug Des.* **100**(1), 136–153 (2022).

Acknowledgements

This work was supported by the State Key Laboratory of Chemical Resources Engineering, Beijing University of Chemical Technology, Beijing 100029, China. The authors extend their appreciation to the Researchers Supporting Project number (RSPD2025R1118) King Saud University, Riyadh, Saud Arabia.

Author contributions

MWI, MS, & IA: Conceptualization, data curation, methodology, software, formal analysis, writing – original draft ZK, & AAM: writing- review, validation. GZ, XS, GAS: supervised the project, QY, XS GAS: supervised the project, funding acquisition, approved and submit the final manuscript.

Funding

The research was financially supported through Researchers Supporting Project Number (RSPD2025R1118), King Saud University, Riyadh, Saudi Arabia.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to A.A.M., X.S. or Q.Y.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.