



Contents lists available at ScienceDirect

## Saudi Pharmaceutical Journal

journal homepage: [www.sciencedirect.com](http://www.sciencedirect.com)

Original article

# Integrated virtual screening, molecular modeling and machine learning approaches revealed potential natural inhibitors for epilepsy

Faez Falah Alshehri

Department of Medical Laboratories, College of Applied Medical Sciences, Ad Dawadimi 17464, Shaqra University, Saudi Arabia

## ARTICLE INFO

## Keywords:

Epilepsy  
S100B  
Machine learning  
Phytochemicals  
Molecular docking

## ABSTRACT

Epilepsy, a prevalent chronic disorder of the central nervous system, is typified by recurrent seizures. Present treatments predominantly offer symptomatic relief by managing seizures, yet fall short of influencing epileptogenesis. This study endeavored to identify novel phytochemicals with potential therapeutic efficacy against S100B, an influential protein in epileptogenesis, through an innovative application of machine learning-enabled virtual screening. Our study incorporated the use of multiple machine learning algorithms, including Support Vector Machine (SVM), k-Nearest Neighbors (kNN), Naive Bayes (NB), and Random Forest (RF). These algorithms were employed not only for virtual screening but also for essential feature extraction and selection, enhancing our ability to distinguish between active and inactive compounds. Among the tested machine learning algorithms, the RF model outshone the rest, delivering an impressive 93.43 % accuracy on both training and test datasets. This robust RF model was leveraged to sift through the library of 9,000 phytochemicals, culminating in the identification of 180 potential inhibitors of S100B. These 180 active compounds were then docked with the active site of S100B proteins. The results of our study highlighted that the 6-(3,12-dihydroxy-4,10,13-trimethyl-7,11-dioxo-2,3,4,5,6,12,14,15,16,17-decahydro-1H cyclopenta[a] phenanthren -17-yl)-2-methyl-3-methylideneheptanoic acid, rhinacanthin K, thiobinupharidine, scopadulcic acid, and maslinic acid form significant interactions within the binding pocket of S100B, resulting in stable complexes. This underscores their potential role as S100B antagonists, thereby presenting novel therapeutic possibilities for epilepsy management. To sum up, this study's deployment of machine learning in conjunction with virtual screening not only has the potential to unearth new epilepsy therapeutics but also underscores the transformative potential of these advanced computational techniques in streamlining and enhancing drug discovery processes.

## 1. Introduction

Epilepsy is a prevalent neurological disorder, characterized by recurrent, unprovoked seizures due to disturbances in the normal pattern of neuronal activity (Blume et al., 2001). This alteration in brain function leads to periods of unusual behavior and sensations, sometimes involving loss of consciousness. Globally, epilepsy affects over 50 million individuals, with the majority (80 %) living in low and middle-income country (Carpio and Hauser 2009). Seizures, the primary symptom of epilepsy, are diverse in their manifestation - ranging from focal onset seizures, where only one area of the brain is affected, to generalized seizures, which involve all areas of the brain (Elshehri et al., 2013). The experience of seizures can include transient confusion, staring spells, uncontrollable jerking movements, and sometimes loss of consciousness or awareness (Josephson et al., 2017). Furthermore,

epilepsy can significantly affect the quality of life of patients, with impacts on mental health, social relationships, and even employment opportunities. Epilepsy's complex etiology involves genetic influences, structural changes in the brain, and functional changes in how neurons behave (Wong 2005). The pathophysiology of epilepsy, or epileptogenesis, is a process where a normal brain is altered to become epileptic due to inciting factors like brain injury, stroke, or prolonged seizures (Pitkänen and Lukasiuk 2009). Despite advances in medical research, no current antiepileptic drugs can inhibit this process of epileptogenesis.

Epilepsy stems from the hyperactivity of neurons, characterized by excessive firing and bursting, as a result of interruptions in the transport of crucial ions like Ca, Na, and K ions through ion channels controlled by voltage and ligands. Furthermore, active phytochemicals that influence  $K^+/Ca^{++}$  channels have been utilized in managing several neurological disorders (Richard 2001). Throughout the progression of epilepsy, there

Peer review under responsibility of King Saud University.

E-mail address: [Falshehri@su.edu.sa](mailto:Falshehri@su.edu.sa).<https://doi.org/10.1016/j.jsps.2023.101835>

Received 5 July 2023; Accepted 18 October 2023

Available online 20 October 2023

1319-0164/© 2023 The Author. Published by Elsevier B.V. on behalf of King Saud University. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

has been noted an overproduction of S100B, a protein that binds to calcium. S100B belongs to S100 protein family, is found within the nucleus and cytoplasm of various cell types and plays a crucial role in controlling a wide range of cellular functions. It plays a significant role in processes like the progression of the cell cycle and differentiation. Furthermore, its link with numerous diseases such as Alzheimer's disease, acute inflammation, cardiomyopathies, rheumatoid arthritis, and cancer has been extensively recorded (Van Eldik and Griffin 1994, Donato 2001). The potential therapeutic benefit of downregulating S100B in epilepsy management has been suggested (Liu et al., 2012).

Computer-assisted drug discovery (CADD) tools have emerged as powerful accelerators in drug discovery processes, bringing down costs considerably (Macalino et al., 2015, Noor et al., 2021). This approach, coupled with the introduction of supercomputing capabilities, novel algorithms, and cutting-edge tools, has significantly elevated the effectiveness of lead discovery in pharmaceutical research (Macalino et al., 2018, Noor et al., 2022). With the integration of artificial intelligence (AI) and machine learning techniques, the analysis of vast data sets related to pharmaceuticals in drug discovery has become more efficient (Floresta et al., 2022, Tahir ul Qamar et al., 2022, Sadaqat et al., 2023). The structure-centric drug development strategy has proved to be particularly useful and effective in identifying and optimizing lead compounds, deepening our molecular understanding of diseases (Yang et al., 2022). In this study, an array of machine learning models was employed to conduct a virtual screening of phytochemicals against the S100B protein, a known drug target in epilepsy. The active hits identified through machine learning were further evaluated using the Lipinski's rule of five, a computational tool to assess their drug-like properties. Phytochemicals demonstrating promising characteristics were further subjected to molecular docking analyses. The results highlighted these phytochemicals as potential inhibitors of the S100B protein, relevant to epilepsy. This research not only opens avenues for the discovery of novel epilepsy therapeutics but also underscores the significant role machine learning can play in expediting and refining drug discovery processes. Nevertheless, *in vitro* validation of these compounds is essential in future studies to elucidate their specific mechanisms of action and validate their potential in addressing this pervasive health challenge.

## 2. Methodology

### 2.1. Preparing and cleaning dataset

The BindingDB database was employed to extract a total of 56 molecules related to the S100B drug target in epilepsy (Sandhu et al., 2022). A further 1801 decoy molecules, also called inactive molecules, were generated utilizing the Database of Useful Decoys (DUDE) (Mysinger et al., 2012). Thus, a collection of 1858 compounds was amassed. From these, 57 molecules sourced from the BindingDB database were tagged as "1" to signify they were active, and conversely, the 1801 decoy compounds were assigned a "0" label indicating their inactive status. While initial dataset was notably imbalanced, having 56 active compounds compared to 1801 inactive ones, careful considerations were made during data splitting. This ensured an equal distribution of both active and inactive compounds in the training and test sets. Such a balanced representation was imperative for the models to achieve a comprehensive understanding of both types of compounds. To further address this inherent imbalance, we utilized the Synthetic Minority Over-sampling Technique (SMOTE). SMOTE helps in creating synthetic samples within the feature space by interpolating between existing instances. This method not only ensures a balanced representation for both classes but also potentially enhances the model's ability to generalize on unseen data.

### 2.2. Data preprocessing and feature extraction

The assembled dataset, consisting of compounds retrieved from the BindingDB database and decoy molecules from the DUDE database, was loaded into a pandas DataFrame in Python (Santos et al., 2020). The compiled dataset was categorized into two primary sections: the features and the target variable. The features were defined by the molecules' SMILES notation, and the target variable was designated to signify the activity status of each molecule, either 'active' labeled as '1' or 'inactive' labeled as '0'. Post this categorization, the dataset was further divided into training and test subsets for model training and validation. This split was accomplished using the `train_test_split` function from the Scikit-learn library (Kramer and Kramer 2016), ensuring an equal distribution of inactive along with active compounds in both sets. The SMILES notation of each molecule was transformed into quantifiable features using the RDKit library (Lovrić et al., 2019). This involved computing 33 features including LogP (lipophilicity), Molecular Weight (MW) and others.

### 2.3. Chemical space and diversity analysis

In addition to the machine learning models and the physicochemical distribution analysis, further investigation into the chemical diversity and similarity among the compounds in the dataset was conducted through a molecular similarity analysis using Tanimoto coefficients. This approach quantifies the degree of similarity between two molecules, based on their molecular fingerprints. First, molecular fingerprints for each compound were computed using the RDKit's Morgan fingerprint algorithm (Bae et al., 2021), capturing the molecular structure information into a binary vector representation. Tanimoto similarity coefficients, metrics that measure the similarity between two molecular fingerprints, were calculated for every pair of molecules in the dataset. These coefficients range from 0, representing completely dissimilar molecules, to 1, signifying identical molecules.

After computing the Tanimoto coefficients for all pairs of molecules, the distribution of these scores was analyzed. Key statistics such as the mean and standard deviation were calculated to understand the overall level and variability of molecular similarity in the dataset. Furthermore, this distribution was visualized using a histogram to get a clearer picture of the diversity in the chemical space of the dataset. This analysis was crucial to ensure the diversity of the dataset and its suitability for training machine learning models. Diverse datasets help prevent overfitting and improve the generalization of the models to unseen data. This comprehensive approach integrating machine learning with chemical space and diversity analysis forms a robust strategy for the development of predictive models in chemoinformatics.

### 2.4. Principle component analysis (PCA)

Next, feature scaling was performed to ensure that all features had a similar scale. This is critical when working with machine learning algorithms that use a distance measure, like K-Nearest Neighbors (KNN). Subsequently, feature extraction was performed using Principal Component Analysis (PCA) to reduce the dimensionality of the data, thereby concentrating the variability of the data into fewer features. PCA a popular technique used for dimensionality reduction and feature extraction, was performed on our data. PCA converts the initial variables into a novel group of variables, referred to as the principal components. These components are uncorrelated and encapsulate the variability observed in the original variables (Prada Gori et al., 2022). In the Scikit-learn implementation (Kramer and Kramer 2016), a PCA object was initialized with the number of components set to 2. This object was fitted on our selected features, and the resulting principal components were used for further processing.

## 2.5. Machine learning models

Several machine learning models were trained on the processed dataset to classify compounds as either active or inactive. The algorithms used for this included Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Naive Bayes (NB), and Random Forest (RF). Each model was tuned and validated using cross-validation and GridSearchCV from the Scikit-learn library.

### 2.5.1. Support Vector Machine (SVM)

SVM is a dynamic and flexible Machine Learning model. It's capable of executing regression, linear/nonlinear classification, and more. It operates by creating a hyperplane in a multidimensional space to differentiate among various classes (Noor et al., 2023). The Scikit-learn's `svm.SVC()` function was used to implement the SVM, with the 'gamma' parameter set to 'scale'. This parameter defines how far the influence of a single training example reaches, with low values meaning 'far' and high values meaning 'close'. We employed 'scale' as it automatically scales the 'gamma' value according to the number of features in the dataset. The kernel parameters being 'linear' and 'rbf' were fed into a grid search for hyperparameter tuning.

### 2.5.2. K-Nearest Neighbors (KNN)

KNN is a non-parametric algorithm that memorizes all extant data points and classifies novel instances according to a specified metric of proximity or similarity (Prada Gori et al., 2022). In this study, the Scikit-learn's `neighbors.KNeighborsClassifier()` function was used to implement the KNN algorithm. The number of neighbors was varied from 1 to 10, and these were fed into a grid search for optimizing the number of neighbors.

### 2.5.3. Naive Bayes (NB)

Naive Bayes classifiers represent a group of basic "probabilistic classifiers" that apply Bayes' theorem, predicated on the substantial assumption of independence among the features (Patel et al., 2020). The Scikit-learn's `naive_bayes.GaussianNB()` function was used to implement the Gaussian Naive Bayes algorithm, with no hyperparameters being tuned since Naive Bayes doesn't typically require this.

### 2.5.4. Random Forest (RF)

Random Forests is a robust machine learning technique that can undertake both classification and regression tasks. It employs numerous decision trees during the training phase, and for classification tasks, it delivers the class that represents the mode of the classes. In the context of regression tasks, it generates the average prediction derived from each individual tree (Breiman 2001). The Scikit-learn's `ensemble.RandomForestClassifier()` function was used to implement the RF. The number of trees in the forest, controlled by the 'n\_estimators' parameter, was set to 100 for our initial model but was subsequently tuned using grid search, with values ranging from 50 to 200. A 'random\_state' parameter was also set to 1 to ensure the reproducibility of our results.

## 2.6. Model evaluation

Each model was trained using the processed dataset and evaluated using 5-fold cross-validation in order to analyze the reliability of result as well as to check the results are not dependent on the specific arrangement of the training data. The hyperparameters of each model were tuned using the GridSearchCV function from Scikit-learn, which performs exhaustive search over specified parameter values for an estimator. Hyperparameters play a pivotal role in defining and refining the performance of machine learning models. Unlike model parameters which are learned during training, hyperparameters are set beforehand. To optimize these for each of our models, we employed the GridSearchCV function from the Scikit-learn library. GridSearchCV methodically searches over a predefined range of hyperparameters,

exhaustively trying out each possible combination. This rigorous process ensures that the best possible set of hyperparameters is chosen, ultimately leading to optimal model performance. The assessment of the models was performed using various metrics, such as recall, precision, F1-score, accuracy, as well as the Area Under the Receiver Operating Characteristic Curve, often abbreviated as AUC-ROC (Ahmad et al., 2021). In addition, the Receiver Operating Characteristic (ROC) curve was constructed, which graphically delineates the effectiveness of the classification model at all possible thresholds. The AUC-ROC, an aggregate measure of the model's performance across all thresholds, was also determined. This metric conveys an overall impression of the model's capacity to discriminate between the different categories, specifically the active and inactive compounds.

## 2.7. Model serialization

Once the model was trained and evaluated, the final model was saved using Python's pickle module for later use. This step, known as model serialization, involved exporting the trained model into a file that could be stored and loaded in the future to make predictions without needing to retrain the model.

## 2.8. Making predictions on a new dataset

Upon the successful serialization of our optimal model, it was deployed to make prognostic evaluations on a novel dataset comprising nearly 9000 phytochemicals of undetermined activity. The library of 9,000 phytochemicals utilized in this research was compiled from various open-source chemical databases, including PubChem (Kim et al., 2019), ChEMBL (Gaulton et al., 2012), and ZINC (Irwin and Shoichet 2005). These databases provide comprehensive information about the structure, properties, and biological activities of small molecules, making them ideal resources for virtual screening and drug discovery studies. The initial dataset's preprocessing and feature extraction strategies were mirrored for this analysis. This preprocessed data was then subjected to the previously trained model, leading to the categorization of each compound into either 'active' or 'inactive'. To refine these results and increase the potential drug-likeness of the selected compounds, we incorporated Lipinski's Rule of Five - a commonly used metric in pharmaceutical research that gauges the likelihood of a chemical compound being an orally active drug in humans (Bashir et al., 2023). This layered approach allowed us to narrow down our list to those phytochemicals which were not only predicted as potentially active but also adhered to the parameters of Lipinski's Rule.

## 2.9. Molecular docking study

### 2.9.1. Preprocessing and validation of target protein

The three-dimensional configuration of the S100B protein, a well-recognized therapeutic target in epilepsy, was obtained from the RCSB Protein Data Bank (Rose et al., 2016). The selected protein structure (PDB ID: [2H61]; Resolution: [1.90 Å]; Organism: [Homo sapiens]; Determination Method: X-ray diffraction) comprised various peptide chains, out of which, Chain A was singled out as the target receptor for our analysis. To begin with, the protein structure was prepared for docking procedures. This involved the removal of undesired water molecules and any linked ligands from the protein structure. Additionally, polar H-atoms were introduced to the structure using the Discovery Studio Visualizer (Systemes 2019).

### 2.9.2. Molecular docking analysis

Using the machine learning model, the phytochemicals classified as active were docked into the S100B protein's active site to facilitate detailed molecular interaction studies. For the docking simulations, we targeted this specific inhibitor binding site. The PyRx tool (Kondapuram et al., 2021), a front-end for AutoDock Vina, was used to conduct these

simulations using a combination of rigid and flexible docking parameters. The docking grid was defined with dimensions of ( $x = 20$ ,  $y = 20$ ,  $z = 20$ ) and positioned to enclose the entire binding site with the coordinates at ( $x = 112.015$ ,  $y = 106.402$ , and  $z = 131.8428$ ). Autodock vina utilized an empirical scoring function to determine the affinity of protein-compound binding, which was calculated by aggregating contributions from various individual terms. The docked complex with the lowest root mean square deviation (RMSD), was considered the optimal complex, and the binding energies among ligand and target protein were evaluated based on their affinity. A good binding strength was indicated by a value  $< -5.00$  kcal/mol, while value  $< -7.00$  kcal/mol indicated very good affinity. These top 5 phytochemicals having highest binding affinity (kcal/mol) were then selected. These top phytochemicals exhibited structural diversity and promising potential as potent inhibitors of the S100B protein, as suggested by their docking scores and interaction patterns.

### 3. Results

#### 3.1. Dataset characteristics and preprocessing

The assembled initial dataset encompassed a total of 1857 compounds. Specifically, 56 molecules known to exhibit activity against the S100B protein target associated with epilepsy were incorporated. The remaining 1801 compounds were decoy molecules. Detailed information pertaining to these 56 active molecules available in [Supplementary File 1: Table S1](#). A thorough assessment of the dataset confirmed its high quality, with no missing values or duplicate entries, hence, qualifying it for further analysis. In the preprocessing stage, each molecule was quantitatively characterized through the transformation of the SMILES notation into numerical descriptors using the RDKit library. A total of 33 features were generated for use in the current study. These features' statistical properties are summarized in [Table 1](#).

**Table 1**  
Descriptive statistics of features extracted from SMILES notation.

Feature	Description	Mean	Standard Deviation	Min	Max
MaxEStateIndex	Maximum electron state indices	12.24435	2.877905	3.449366	15.89941
MinEStateIndex	Minimum electron state indices	-1.65521	2.126304	-8.89218	1.049444
MaxAbsEStateIndex	Maximum absolute electron state indices	12.24435	2.877905	3.449366	15.89941
MinAbsEStateIndex	Minimum absolute electron state indices	0.143983	0.175193	0.000205	1.564815
qed	Quantitative Estimation of Drug-likeness	0.336721	0.255879	0.026506	0.917744
MolWt	Molecular weight	531.7295	180.852	166.288	884.158
HeavyAtomMolWt	Molecular weight of heavy atoms	501.5868	174.5467	146.128	858.161
ExactMolWt	Exact molecular weight	531.0794	180.6201	166.159	883.4894
NumValenceElectrons	Number of valence electrons	193.3242	65.53219	64	338
MaxPartialCharge	Maximum partial charge	0.290576	0.09857	0.036967	0.572618
MinPartialCharge	Minimum partial charge	-0.44589	0.103825	-0.87236	-0.19716
MaxAbsPartialCharge	Maximum absolute partial charge	0.454735	0.098555	0.240389	0.87236
MinAbsPartialCharge	Minimum absolute partial charge	0.281731	0.088753	0.036967	0.467962
BalabanJ	Balaban topological index	1.745196	0.608013	0.686564	6.220887
BertzCT	Bertz molecular complexity index	1285.799	611.2487	119.2587	3136.19
MolLogP	Partition coefficient between octanol and water	5.183181	2.522943	-4.0758	15.0614
MolMR	Molecular molar refractivity	138.8478	48.9564	25.7661	259.0877
HeavyAtomCount	Number of heavy atoms	36.58365	12.75426	11	66
NHOHCount	Number of hydroxyl and amine groups	2.206802	1.224305	1	8
NumHDonors	Number of hydrogen bond donors	1.860121	1.011279	0	8
NumHAcceptors	Number of hydrogen bond acceptors	5.878223	3.123946	0	13
NumRotatableBonds	Number of rotatable bonds	8.279759	4.326629	0	23
NumHeteroatoms	Number of non-carbon atoms	9.281953	4.797838	1	34
NumAromaticRings	Number of aromatic rings	2.877126	1.844158	0	8
NumSaturatedRings	Number of saturated rings	0.686231	1.171966	0	8
NumAliphaticRings	Number of aliphatic rings	1.234229	1.291062	0	8
NumAromaticHeterocycles	Number of aromatic heterocyclic rings	0.809106	0.885165	0	5
NumSaturatedHeterocycles	Number of saturated heterocyclic rings	0.312671	0.771962	0	6
NumAliphaticHeterocycles	Number of aliphatic heterocyclic rings	0.729018	0.935967	0	6
NumAromaticCarbocycles	Number of aromatic carbocyclic rings	2.06802	1.491941	0	6
NumSaturatedCarbocycles	Number of saturated carbocyclic rings	0.37356	0.844887	0	8
NumAliphaticCarbocycles	Number of aliphatic carbocyclic rings	0.505211	0.942269	0	8
FractionCSP3	Fraction of carbons that are sp <sup>3</sup> hybridized	0.404419	0.262591	0	1

Post preprocessing, the dataset was split into training and test sets. Both of these sets were judiciously represented with active and inactive compounds. The compiled datasets for these training and test sets are respectively detailed in [Supplementary File 2: Table S1](#) and [Supplementary File 3: Table S1](#). Within these [supplementary files](#), the SMILES notation, binary activity labels, and all extracted features for each molecule are documented, thereby providing complete transparency of our computational workflow and dataset construction.

#### 3.2. Principle component analysis

In our study, we employed Principal Component Analysis (PCA) to transform the original 33 descriptors, which represent the distinct characteristics of the compounds, into two principal components. These components were observed to capture a significant proportion of the variance in our dataset. The eigenvalues, a measure of the variance explained by each principal component, were  $3.27111643e + 04$  for the first component and  $2.64810671e + 00$  for the second component. These values indicate the amount of variance each component accounts for in the dataset. The larger the eigenvalue, the more variance (information) that component captures from the dataset. In our case, the first principal component, with a significantly larger eigenvalue, explained approximately 46.60 % of the variance, encapsulating nearly half of the critical information embedded in our original high-dimensional data. The second component contributed an additional 10.17 % to the explained variance. The eigenvalues illustrate that these two principal components effectively retain a substantial proportion of the essential information from the dataset. This highlights the power of PCA as a dimensionality reduction technique, especially beneficial in managing high-dimensional data in our study.

In [Fig. 1](#), a scatter plot of these two principal components displays a distinct separation between the active and inactive phytochemical compounds. This visual differentiation reveals that the PCA-derived



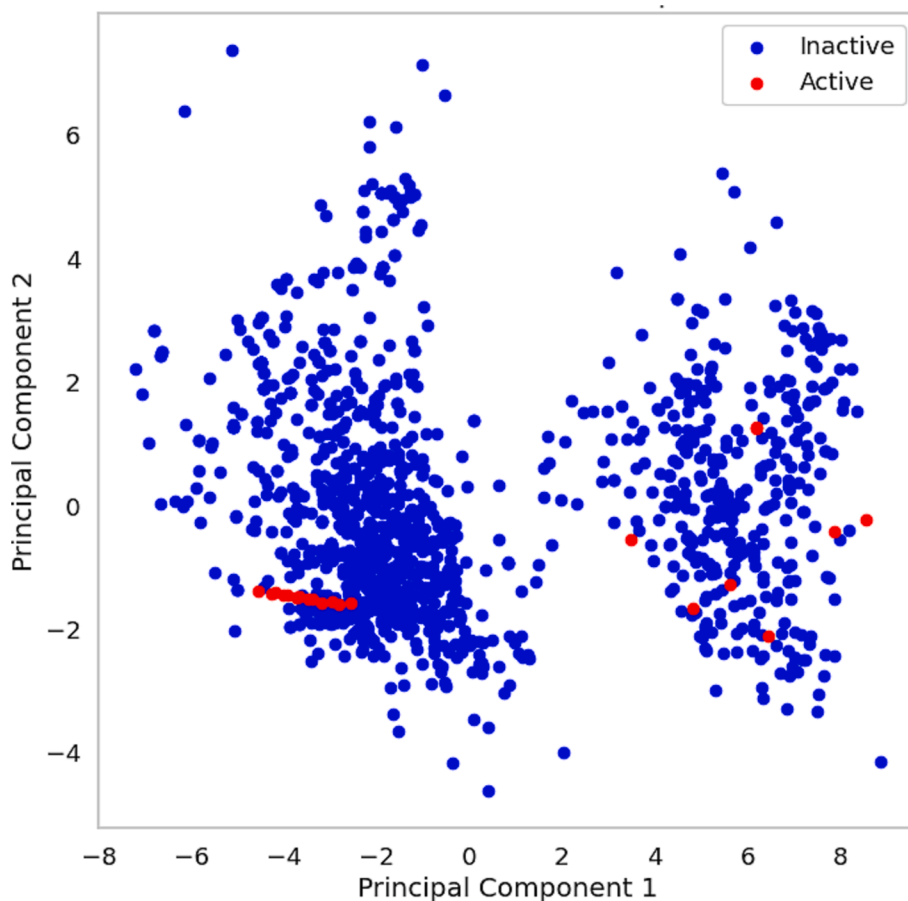


Fig. 1. Scatter plot of the two principal components derived from PCA. Active compounds are marked in red, and inactive compounds are marked in blue.

components can serve as robust discriminative features to distinguish between the two classes of compounds. The variance explained by these components, along with the clear classification in the scatter plot, reinforces the utility of PCA in unearthing critical, non-redundant information in high-dimensional data. This transformation not only simplified the complexity of our data but also facilitated a more efficient and streamlined analysis. The successful reduction of dimensions using

PCA will enable the construction of subsequent machine learning models with enhanced interpretability and potentially improved performance, particularly in predicting the activity of phytochemical compounds. Our findings illuminate the promise of PCA in extracting meaningful insights from high-dimensional chemical descriptor data, paving the way for future studies in this exciting intersection of chemistry and data science.

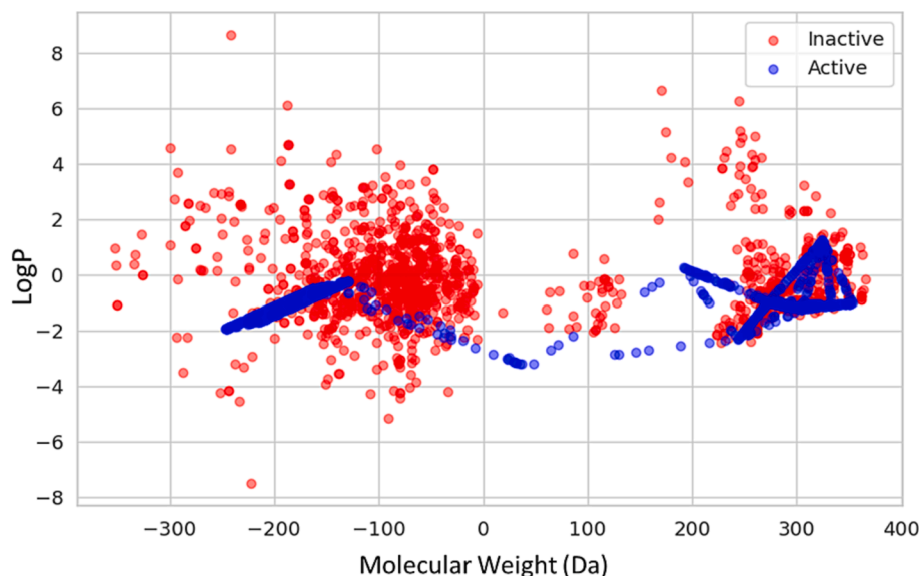


Fig. 2. The distribution of chemical space and diversity within the training set. The parameters defining the chemical space include the LogP molecular weight.

### 3.3. Chemical space and diversity analysis

The efficiency of machine learning models is predominantly affected by the chemical heterogeneity of the samples in both the training and test datasets. Models trained on a diverse set of samples are more likely to generalize well to unseen data. In framework of current study, we conducted a physicochemical distribution analysis of the training and test sets with respect to the two principal components: molecular weight (MW) and LogP. The MW in our dataset ranges from 50 to 500 Da (Daltons) and LogP varies from  $-2$  to  $15$ .

The analysis of the chemical space occupied by the training set (Fig. 2) reveals that active compounds, represented in blue, tend to cluster in regions with higher values of the first principal component, suggesting a higher molecular weight. Inactive compounds, represented in red, spread more evenly across the first principal component, indicating a wider range of molecular weights. The distribution of LogP values, represented by the second principal component, is also varied within each group, indicating a wide range of lipophilicity among the compounds. The test set (Fig. 3) demonstrates a similar distribution, validating the representativeness of our training set and ensuring that our model is evaluated on a test set that shares the same chemical space as the training data.

We also performed a molecular similarity analysis by calculating the Tanimoto coefficients for pairs of molecules in our dataset (Fig. 4). Tanimoto coefficient is a popular metric in chemoinformatics for quantifying the similarity between two molecules based on their molecular fingerprints. The coefficients range from 0 (no similarity) to 1 (identical molecules). Our analysis yielded a mean Tanimoto score of  $\sim 0.12$ , suggesting that, on average, the molecules in our dataset share about 12 % of their features. This implies a moderate level of similarity among the molecules in our dataset. The standard deviation of  $\sim 0.07$  reveals a significant variability in molecular similarity within our dataset, implying a considerable diversity among the molecules. This diversity is beneficial for training robust machine learning models as it allows the models to learn and capture a wide range of molecular characteristics.

The compounds (inactive and active) in both the training and test sets exhibit a wide range of values for the two principal components (Table 2). The mean values indicate that active compounds in the training set generally have higher molecular weight and are less lipophilic than inactive compounds. However, the standard deviation shows substantial variability within each group. These findings underline the

complex relationship between the molecular properties of compounds and their biological activity, necessitating the use of sophisticated machine learning techniques for accurate prediction.

### 3.4. Model generation and validation

To categorize the active inhibitors against S100B, our study utilized various machine learning algorithms, namely kNN, SVM, RF, and NB. These models were developed utilizing the Python sklearn library and trained on a dataset extracted from the Binding DB database. The effectiveness of these models was evaluated using various statistical metrics, including accuracy, recall (sensitivity), specificity, MCC, and the AUC. The performance of each model on the test set is demonstrated in Table 3.

The results demonstrate that the RF model displayed superior performance across all metrics. It achieved a specificity of 0.947653, sensitivity of 0.920152, accuracy of 0.934259, MCC of 0.868600, and AUC of 0.980117 (Supplementary File 4). This indicates that the RF model provides an excellent balance between predicting true positives (active compounds) and true negatives (inactive compounds), thus making it the most reliable model among the ones tested.

Meanwhile, the kNN model also performed well with a fairly high accuracy of 0.851852 and a satisfactory MCC of 0.705018, indicating its utility as a good secondary model for this classification task. The SVM and GNB models, on the other hand, demonstrated a high sensitivity but lagged in terms of specificity and MCC. These models, while good at identifying true positives, have a higher rate of false positives. Therefore, while these models can be useful in contexts where missing a positive case would be detrimental, they are not as efficient in general classification tasks for this specific dataset. In comparison to other employed machine learning models, the RF model emerged as superior in terms of both accuracy and the MCC. It is worth-noting that the performance of a model is directly correlated with the AUC. Notably, the RF model displayed the highest AUC, trailed by the SVM model, as evidenced in Fig. 5 depicting performance on both the training and test sets.

Following that, the RF model was employed to classify the active phytochemicals that are effective against S100B. Remarkably, from library of 9000 active compounds a total of 584 phytochemicals were predicted to be active for S100B protein. This highlights the utility and accuracy of the RF model in predicting active compounds in this context.

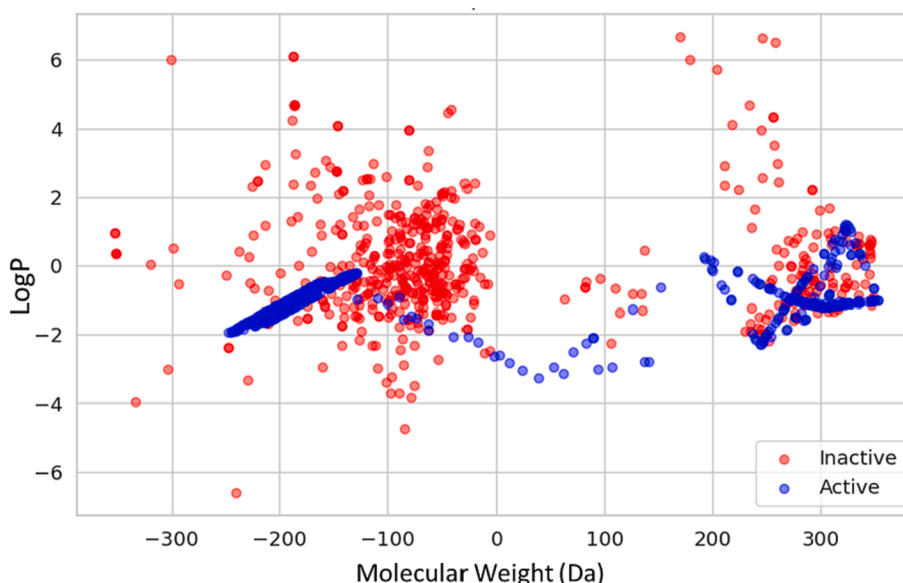
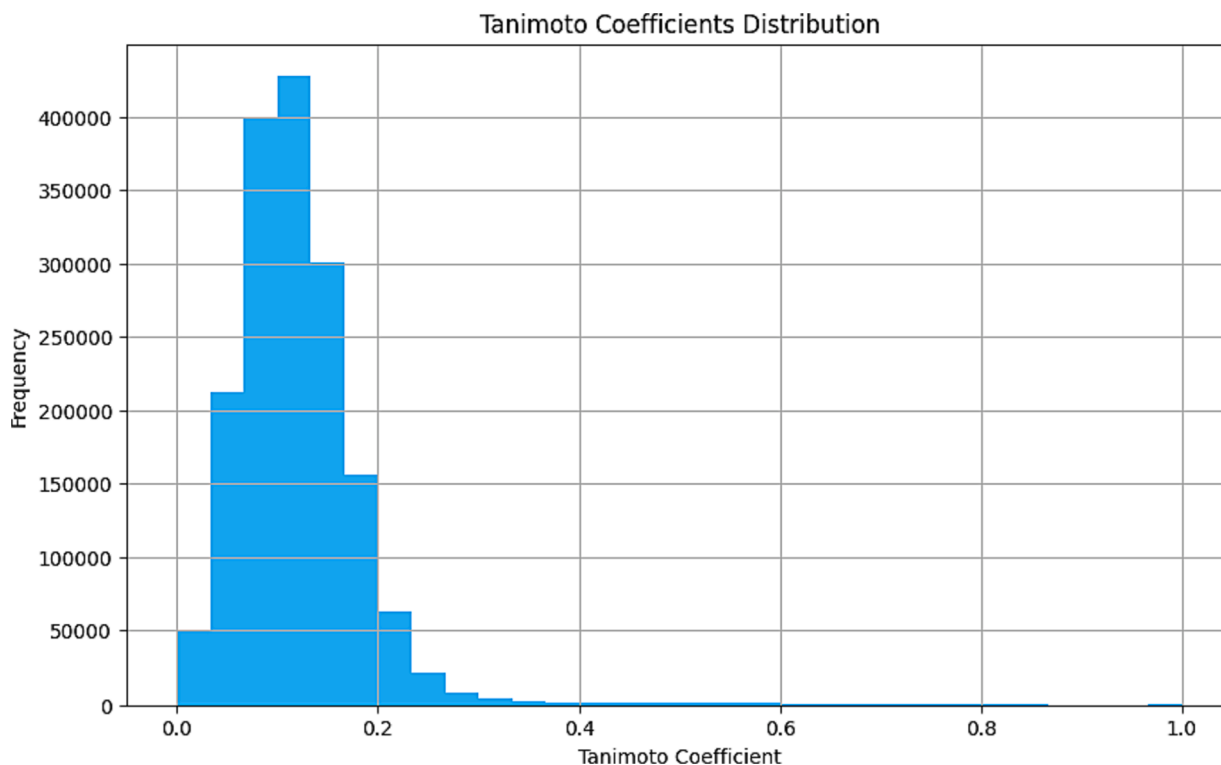


Fig. 3. The distribution of chemical space and diversity within the test set. The parameters defining the chemical space include the LogP molecular weight.



**Fig. 4.** Histogram of the Tanimoto scores provides a visual representation of the distribution of pairwise molecular similarities in our dataset. The broad spread of the histogram indicates a diverse chemical space, which is essential for avoiding overfitting in machine learning models.

**Table 2**

Summary statistics for the principal components.

Datasets	Statistics	Principal Component 1	Principal Component 2
Train Active	Mean	13.23591	-1.00004
	Standard Deviation	234.9012	0.646442
	Min	-246.599	-3.20646
	Max	353.4598	1.25585
Train Inactive	Mean	4.30182	-0.04584
	Standard Deviation	183.2961	1.592324
	Min	-352.438	-7.50709
	Max	365.4606	8.652066
Test Active	Mean	-1.62929	-1.04072
	Standard Deviation	230.2291	0.651726
	Min	-247.196	-3.24771
	Max	352.6287	1.19385
Test Inactive	Mean	-9.94967	0.147748
	Standard Deviation	172.438	1.706272
	Min	-352.438	-6.61844
	Max	349.4086	6.666681

**Table 3**

Performance evaluation metrics of predicted models.

Model	Accuracy	Sensitivity	Specificity	MCC	AUC
kNN	0.851852	0.876426	0.82852	0.705018	0.904505
SVM	0.769444	0.918251	0.628159	0.568288	0.842096
RF	0.934259	0.920152	0.947653	0.8686	0.980117
NB	0.731481	0.914449	0.557762	0.502821	0.827147

### 3.5. Analyzing drug-like potential of active compounds

Following the determination of the active phytochemicals, we proceeded to analyze their drug-likeness, a vital parameter in evaluating potential therapeutic agents. Utilizing RDKit's Lipinski module, we computed key molecular properties including MW, HBD, HBA, and the LogP for these active compounds. Our drug-likeness criteria were established based on Lipinski's Rule of Five, which asserts that a compound is likely to have favorable absorption or permeation characteristics if it possesses a MW < 500 Daltons, < 5H-bonds donors, < 10H-bonds acceptors, and a LogP value < five. Remarkably, out of the 584 active phytochemicals, 180 fulfilled these criteria, demonstrating significant potential for further docking studies and exploration as drug candidates. To comprehend the relationships and correlation among these molecular properties, we generated a heatmap of the correlation matrix (Fig. 6). This heatmap provides a visual representation of the correlation coefficients between each pair of properties (MW, HBD, HBA, LogP). In this heatmap, the strength and nature of correlations are visualized with varying color intensities, ranging from deep red (indicating a strong negative correlation) to deep blue (indicating a strong positive correlation). A value closer to zero, such as the correlation between HBA and MW at -0.013, is visualized in a neutral color.

To visualize the distribution of these drug-like properties among the selected compounds, we generated a scatter matrix plot (Fig. 7). This plot, demonstrated by the seaborn pairplot function, provides a pairwise relationship for an array of variables. Each data point represents a specific compound, plotted in a multidimensional space of molecular properties. The kernel density estimation (KDE) on the diagonal allows us to see the distribution of values for each property.

Overall, the scatter matrix reveals notable clusters, suggesting that our selection of potential drug-like phytochemicals shares similar molecular properties, which is promising for further pharmaceutical development. These observations, combined with the machine learning prediction results, provide a compelling foundation for future experimental validation and potential therapeutic applications.

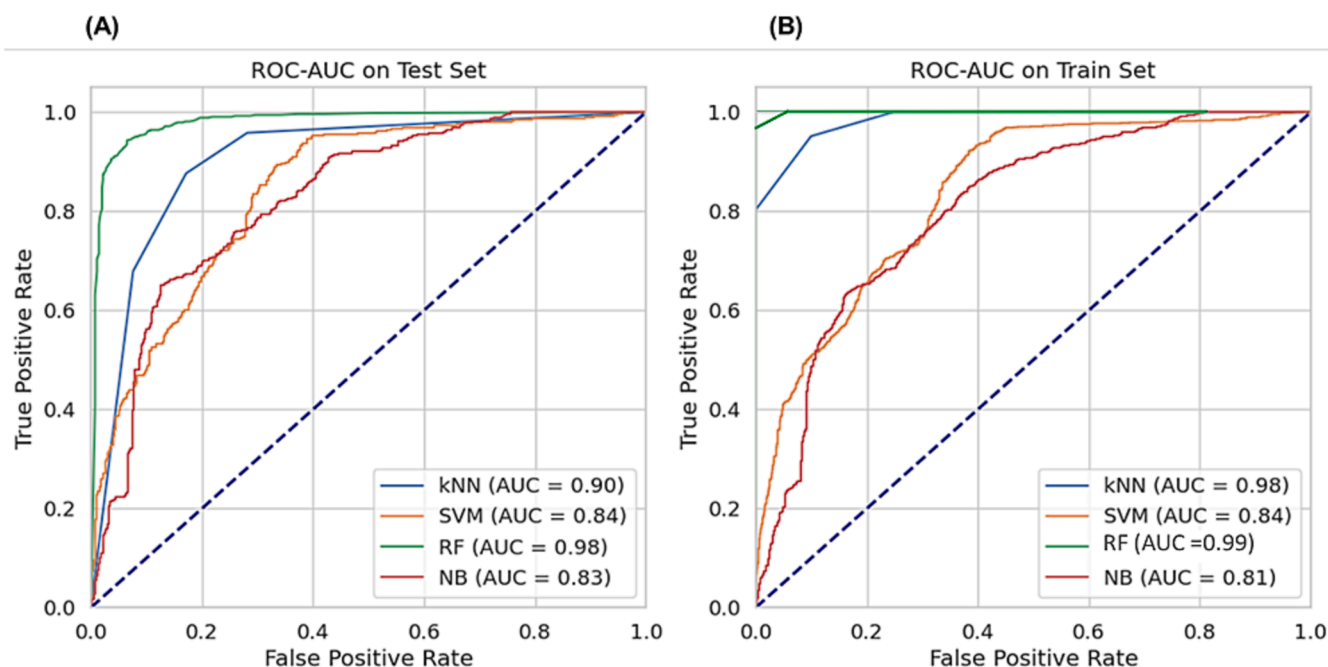


Fig. 5. The ROC-AUC curve of all the models on (A) Test set (B) Train set. The graph shows the TP against FP rate.

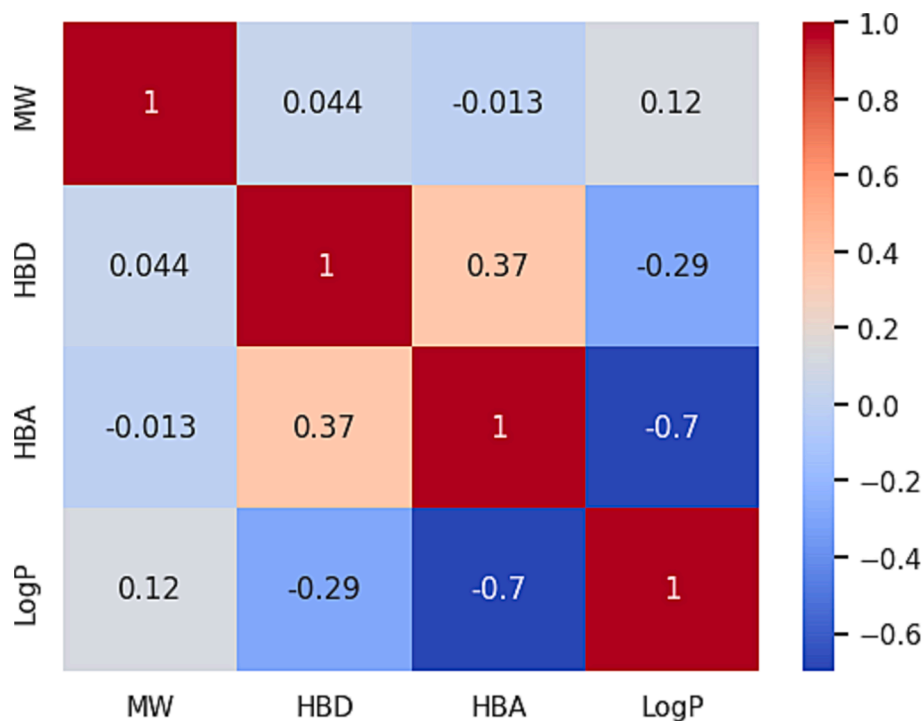


Fig. 6. Heatmap of the correlation matrix for molecular properties. The heatmap provides a visual representation of the correlation coefficients between each pair of properties (MW, HBD, HBA, LogP). Darker colors indicate stronger correlations.

### 3.6. Molecular docking analysis

In our rigorous investigation, we executed an extensive molecular docking analysis with the S100B protein utilizing a sophisticated machine learning-based virtual screening technique. A comprehensive library of 180 phytochemicals was subjected to this screening, and the quintet of compounds displaying the most robust binding affinity were earmarked for further scrutiny. The phytochemical with the highest binding affinity was the complex compound 6-(3,12-dihydroxy-4,10,13-

trimethyl-7,11-dioxo-2,3,4,5,6,12,14,15,16,17-decahydro-1H-cyclopenta[a] phenanthren-17-yl)-2-methyl-3-methylideneheptanoic acid. This compound exhibited a binding affinity of  $-8.55412$  and an RMSD (Root Mean Square Deviation) value of  $1.538235 \text{ \AA}$  (Table 4). Its potent interaction was majorly with Glu F:58 and Lys F:29 residues. Next Rhinacanthin K, showing a binding affinity of  $-8.13617$  and an RMSD of  $3.478616 \text{ \AA}$ . The residues involved in its interplay with the S100B protein were Lys F:29 and Asp F:61. Thiobinupharidine was the third-highest binder, possessing a binding affinity of  $-8.09721$  and RMSD



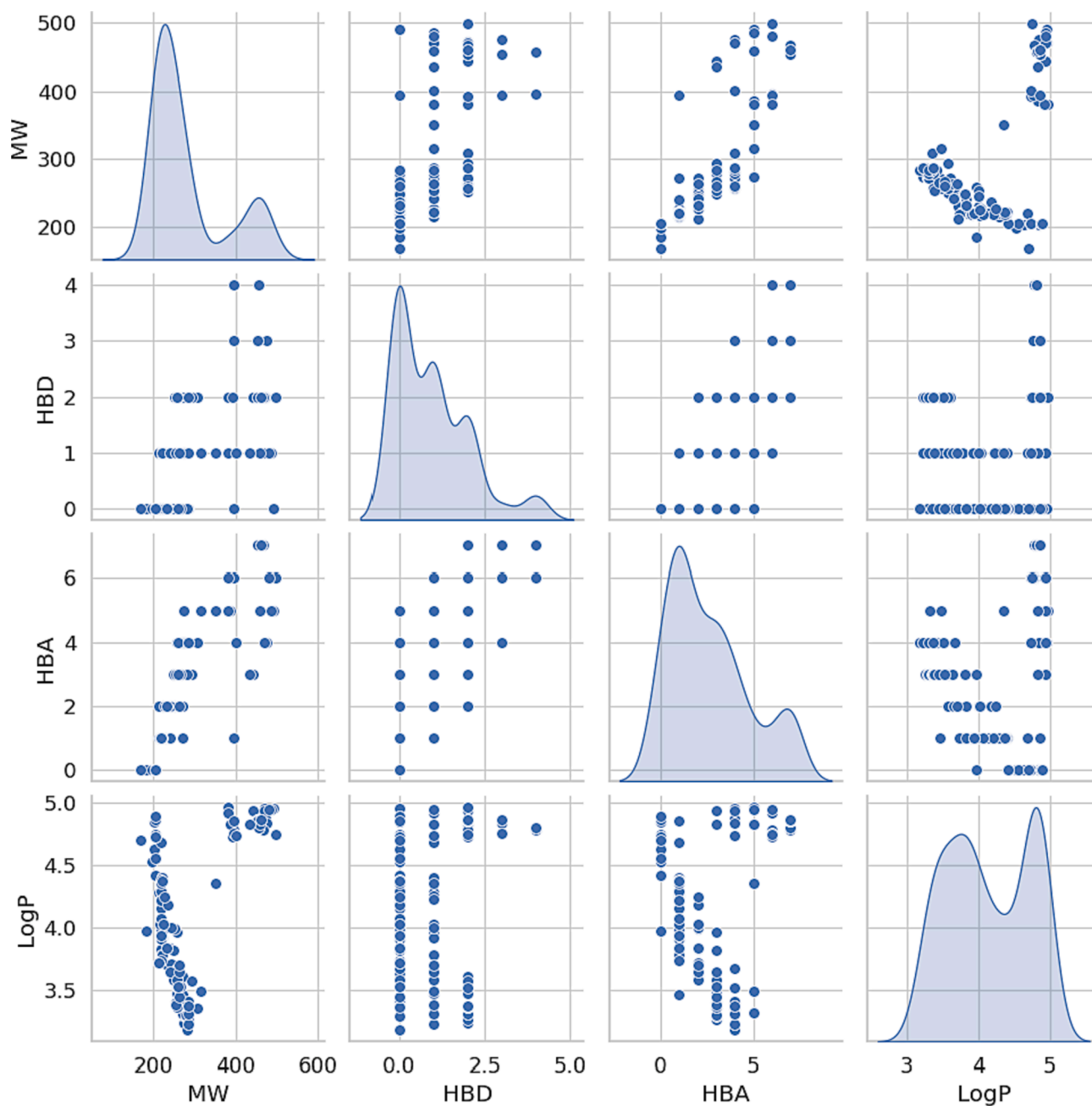


Fig. 7. Scatter matrix of molecular properties. This scatter matrix presents a pairwise comparison of molecular properties (MW, HBD, HBA, LogP) for the selected phytochemicals. The diagonal plots indicate the kernel density estimate of each property.

value of 3.133208 Å, engaging principally with Asp F:61 residue. Scopadulcic Acid, the fourth compound, displayed a binding affinity of  $-7.75486$  and an RMSD of 2.324887 Å, predominantly interacting with the Asp F:62 residue. Lastly, Maslinic Acid, with a binding affinity of  $-7.29647$  and RMSD of 3.182368 Å, was found to interact with Gly F:64 residue (Fig. 8). In conclusion, this intricate docking analysis elucidated specific interactions of these phytochemicals with the S100B protein, unearthing their potential as novel therapeutic candidates.

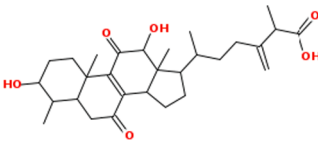
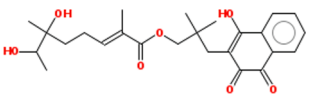
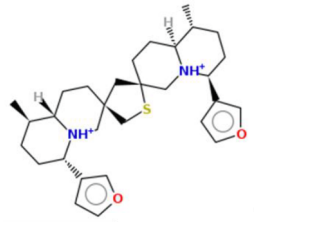
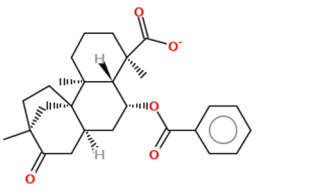
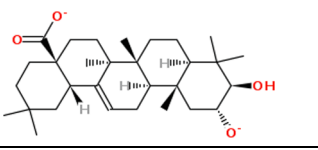
#### 4. Discussion

Epilepsy, a chronic neurological disorder, is characterized by recurring seizures, and it affects millions of individuals worldwide (Hu et al., 2023). The etiology of epilepsy is multifaceted, encompassing genetic mutations, brain injuries, and imbalances in neurotransmitter

function (Helbig et al., 2017). Despite significant advances in understanding the disorder's pathophysiology, effective treatment options remain limited, particularly for drug-resistant epilepsy. Therefore, the development of novel, effective therapeutic strategies for epilepsy is crucial. S100B, a protein predominantly expressed in astrocytes, plays a significant role in epilepsy. Abnormally high levels of S100B have been detected in the cerebrospinal fluid and serum of patients with epilepsy, indicating its potential as a therapeutic target (Liang et al., 2019, Langeh and Singh 2021). The functional relevance of S100B in epilepsy, combined with the absence of closely related human homologs, makes S100B an attractive target for the development of antiepileptic medications.

Traditionally, the development of new drugs targeting specific proteins such as S100B has been a lengthy and costly endeavor. However, the advances in CADD, now offer a rapid and precise methodology for

**Table 4**  
Binding affinity and RMSD values of docked complexes.

PubChem ID	Phytochemical name	Binding affinity (kcal/mol)	RMSD (Å)	2D structures
10,838,646	6-(3,12-dihydroxy-4,10,13-trimethyl-7,11-dioxo-2,3,4,5,6,12,14,15,16,17-decahydro-1H-cyclopenta[a]phenanthren-17-yl)-2-methyl-3-methylideneheptanoic acid	-8.55412	1.538235	
10,765,714	Rhinacanthin K	-8.13617	3.478616	
442,554	Thiobinupharidine	-8.09721	3.133208	
11,729,855	Scopadulcic Acid	-7.75486	2.324887	
73,659	Maslinic Acid	-7.29647	3.182368	

screening extensive libraries of active phytomolecules. This approach holds the potential to hasten the discovery and development of novel therapeutic agents targeting S100B, thus significantly influencing epilepsy treatment. The incorporation of machine learning techniques has further revolutionized this approach, by offering an efficient method for classifying prospective active and inactive compounds against protein targets. This study intends to harness the power of these machine learning algorithms to enhance the drug discovery process, focusing particularly on the identification of novel inhibitors against the S100B protein. The emphasis on machine learning not only aids in improving the precision of virtual screening but also aids in reducing the frequency of false-positive hits. This dual advantage significantly boosts the overall efficiency and accuracy of drug discovery endeavors. As a result, machine learning brings a new level of sophistication and capability to the drug discovery process, reinforcing its importance in modern medicinal chemistry.

A comprehensive dataset of 1857 compounds, inclusive of 56 known active agents against S100B and 1801 decoy molecules was used in current study. Following the identification of active and inactive compounds, it was imperative to comprehend the characteristics or features that make certain compounds effective against the S100B protein in treating epilepsy. For this, we implemented a feature extraction and

selection process, including the analysis of physicochemical properties, topological descriptors, and structural fingerprints. These features, essential in drug discovery, provide insights into the compound's reactivity, stability, and interactions with the target protein.

PCA was performed to reduce the complexity of our multidimensional dataset while preserving its variance. This process yielded key insights into the main features contributing to compound activity. Following PCA, we constructed machine learning models using KNN, SVM, and RF algorithms, employing the selected features as input. Each model was trained and validated on the curated dataset, with the aim to distinguish active from inactive compounds. These models' performances were assessed based on accuracy, precision, recall, F1-score, and AUC-ROC.

Among the three, the RF model outperformed others in predicting active compounds, which could be due to its capability to handle high dimensional data, consider interaction effects, and mitigate overfitting. Notably, RF model also provides feature importance, offering insight into which features were most influential in classifying a compound as active or inactive. This feature relevance information is pivotal in understanding the characteristics crucial for a compound's effectiveness against S100B protein.

After establishing the RF model as the most reliable one, it was used

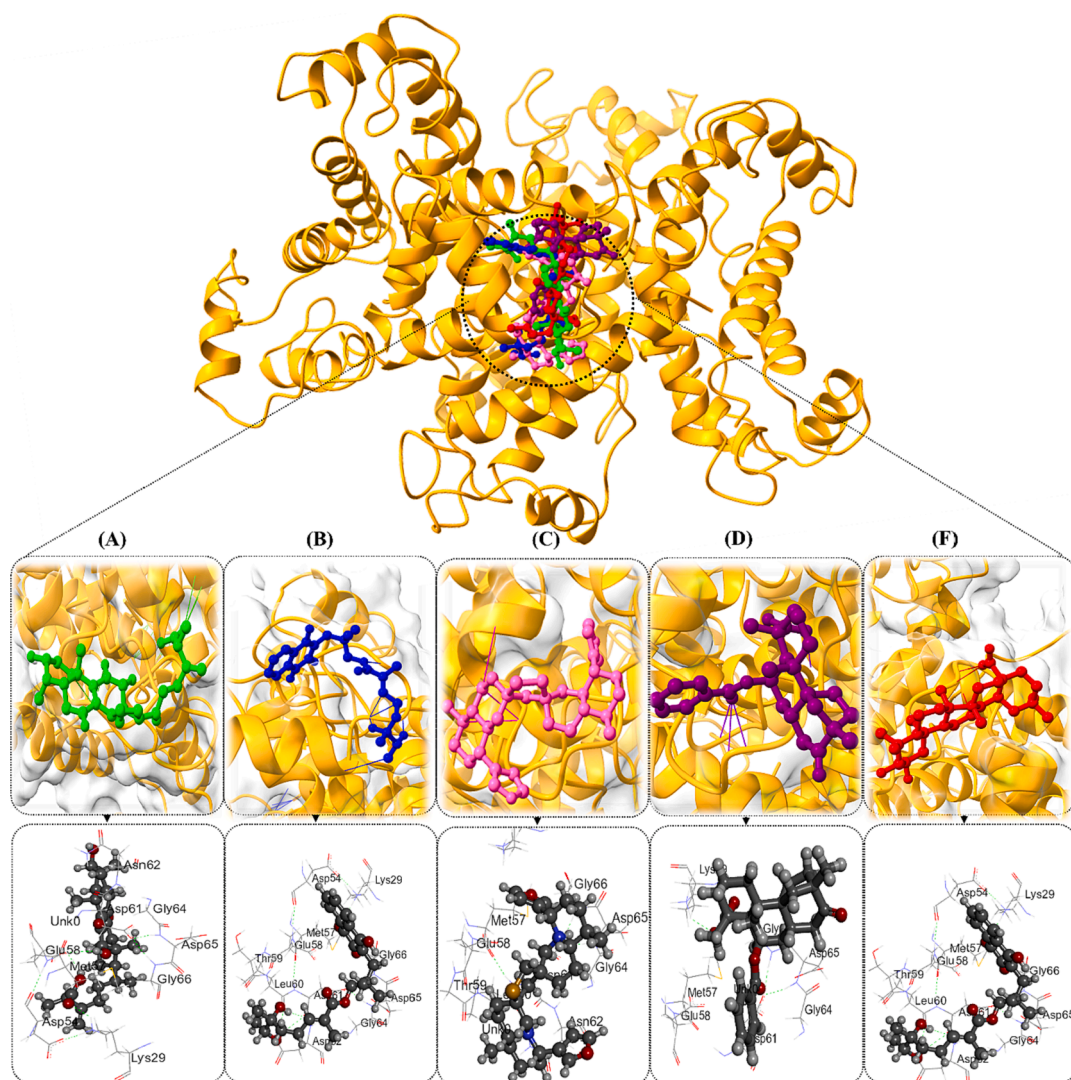


Fig. 8. 3D visualization of active compounds with target proteins.

to screen a library of 9000 phytochemicals, revealing 584 compounds predicted as active against the S100B protein. This provided a broad pool of potential candidates for drug development in the treatment of epilepsy. Further examination of these 584 compounds was undertaken with a specific focus on their drug-likeness, using the Lipinski's Rule of Five. This rule is a well-accepted guideline in the field of drug discovery, utilized to predict the oral bioavailability of a compound. Upon the utilization of Lipinski's Rule, we narrowed down our list to 180 compounds that met these criteria and thus, were likely to be well-absorbed in the human body.

Molecular docking was performed with these 180 compounds to simulate their interactions with the S100B protein. Our study identified five with superior binding affinity towards the S100B protein, which could potentially contribute to the development of novel therapeutics for epilepsy. The phytochemical demonstrating the most potent binding affinity was the elaborately structured 6-(3,12-dihydroxy-4,10,13-trimethyl-7,11-dioxo-2,3,4,5,6,12,14,15,16,17-decahydro-1H-cyclopenta [a]phenanthren-17-yl)-2-methyl-3-methylideneheptanoic acid. Rhinacanthin K, a well-known compound with documented medicinal benefits, was a close contender. Thiobinupharidine, Scopadulcic Acid, and Maslinic Acid were also among the top performing compounds.

Samad et al. (Samad et al., 2023) integrated machine learning with virtual screening and unveiled potential inhibitor against 3CL<sup>PTO</sup> of SARS-CoV-2. To sum up, our study has considerable implications for the

field of drug discovery, particularly in terms of finding therapeutics for epilepsy. Epilepsy has remained a complex condition to manage, and current treatments only aim at managing symptoms without fundamentally altering the disease course. Our study's utilization of a machine learning-based approach has allowed us to screen an extensive library of phytochemicals and identify potential lead compounds that interact with the S100B protein, a critical target for epilepsy. This targeted approach could potentially lead to the discovery of drugs that can impact the disease's underlying mechanisms, rather than just controlling the symptoms. Furthermore, this study can streamline the drug discovery process, which traditionally has been time-consuming and resource-intensive, by integrating computational methods to pinpoint potential candidates for further study.

Our study leveraged machine learning and virtual screening to identify several compounds with demonstrated efficacy in epilepsy management. The diverse pharmacological properties of compounds predicted in current study, including antioxidant and anti-inflammatory effects, emphasize their potential for future research to explore its impact on neurological conditions. While it's speculative at this point, compounds with antioxidant and anti-inflammatory properties could play a role in neuroprotection or modulation of neural pathways, which could have implications for conditions like epilepsy. As with any potential treatment, rigorous scientific investigation would be essential to establish efficacy, safety, and appropriate application.

Additionally, this also has some limitation, as for instance one major constraint is the reliance on the quality and diversity of our data set. Although we used a broad phytochemical library, the effectiveness of our machine learning models is dependent on the training dataset's quality and diversity. Therefore, any bias or lack of diversity may limit the prediction outcomes. Moreover, the interactions identified via *in silico* methods are theoretical and must be validated through *in vivo* and *in vitro* experimental studies. These findings need to be further evaluated in cell-based assays and animal models to establish their therapeutic relevance. Also, pharmacokinetic and pharmacodynamic studies are necessary to ensure the efficacy and safety of the identified compounds in humans. Moving forward, the future prospects of this research involve expanding the phytochemical library and incorporating more diverse datasets for training and testing our machine learning models. Additionally, improving our understanding of the molecular mechanisms underlying the interactions between the identified phytochemicals and the S100B protein could provide valuable insights into the design of novel drugs for epilepsy. Further experimental validation, including *in vitro* and *in vivo* studies, is critical to affirm the therapeutic potential of these compounds. Finally, the findings from this study could stimulate similar research approaches for other diseases, thereby promoting the integration of machine learning in drug discovery pipelines across various therapeutic areas.

While our study adopted a sequential approach for drug design, focusing on potency first and then other drug-like properties, it's essential to note that drug design can benefit from a multi-objective approach. By simultaneously considering multiple properties, the drug discovery process can potentially be streamlined, improving efficiency and reducing costs associated with trial and error. Such an approach can also increase the chances of identifying trustworthy candidates from the onset. In our machine learning-based virtual screening, the selection of 33 features, including molecular weight, MlogP, and counts of hydrogen bond acceptors and donors, demonstrates our detailed and comprehensive approach to identifying potential candidates. In the context of our study, while the sequential approach yielded significant findings, considering drug-likeness guidelines such as Lipinski's rule of five early in the drug design phase might be advantageous.

## 5. Conclusion

In conclusion, the present study has made significant strides towards advancing our understanding of S100B protein as a potential therapeutic target in epilepsy. Leveraging the potential of machine learning algorithms and computer-aided drug design, our study have expedited the traditionally laborious and time-intensive process of drug discovery. Utilizing a comprehensive dataset of compounds, we developed and validated machine learning models that efficiently differentiated between active and inactive compounds against the S100B protein. Among the employed algorithms, the Random Forest model outperformed others, demonstrating high predictive accuracy and providing valuable insights into feature importance. Subsequent application of this model enabled the screening of a large library of phytochemicals, culminating in the identification of 584 compounds projected to be active against S100B. The application of Lipinski's Rule of Five further refined this list to 180 compounds with desirable drug-likeness attributes. Molecular docking studies of these 180 compounds revealed five phytochemicals with superior binding affinity towards the S100B protein, thus highlighting their potential as leads for anti-epileptic drug development. Overall, our work shows how combining machine learning with drug design can speed up drug discovery. While our findings are promising, we understand the need for a broader validation of the proposed CAAD workflow. As such, we have accentuated in our discussion the importance of future studies spanning different datasets and conditions. Though our focus was primarily on epilepsy and the S100B protein, we believe that our methodology holds potential for other ailments and targets. Such endeavors could catalyze the creation of specialized and

efficient treatments, promising better care for countless epilepsy patients worldwide.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgement

The author would like to thank the Deanship of Scientific Research at Shaqra University for supporting this work.

## Appendix A. Supplementary material

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.jsps.2023.101835>.

## References

- Ahmad, A., Akbar, S., Khan, S., et al., 2021. Deep-AntiFP: Prediction of antifungal peptides using distant multi-informative features incorporating with deep neural networks. *Chemom. Intel. Lab. Syst.* 208, 104214.
- Bae, S.-Y., Lee, J., Jeong, J., et al., 2021. Effective data-balancing methods for class-imbalanced genotoxicity datasets using machine learning algorithms and molecular fingerprints. *Comput. Toxicol.* 20, 100178.
- Bashir, Y., Noor, F., Ahmad, S., et al., 2023. Integrated virtual screening and molecular dynamics simulation approaches revealed potential natural inhibitors for DNMT1 as therapeutic solution for triple negative breast cancer. *J. Biomol. Struct. Dyn.* 1–11.
- Blume, W.T., Lüders, H.O., Mizrahi, E., et al., 2001. Glossary of descriptive terminology for ictal semiology: report of the ILAE task force on classification and terminology. *Epilepsia* 42, 1212–1218.
- Breiman, L., 2001. Random Forests. *Machine Learn.* 45, 5–32.
- Carpio, A., Hauser, W.A., 2009. Epilepsy in the developing world. *Curr. Neurol. Neurosci. Rep.* 9, 319.
- Donato, R., 2001. S100: a multigenic family of calcium-modulated proteins of the EF-hand type with intracellular and extracellular functional roles. *Int. J. Biochem. Cell Biol.* 33, 637–668.
- Elshoff, L., Muthuraman, M., Anwar, A.R., et al., 2013. Dynamic imaging of coherent sources reveals different network connectivity underlying the generation and perpetuation of epileptic seizures. *PLoS One* 8, e78422.
- Floresta, G., Zagni, C., Gentile, D., et al., 2022. Artificial Intelligence Technologies for COVID-19 De Novo Drug Design. *Int. J. Mol. Sci.* 23, 3261.
- Gaulton, A., Bellis, L.J., Bento, A.P., et al., 2012. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* 40, D1100–D1107.
- Helbig, I., von Deimling, M., Marsh, E.D., 2017. Epileptic encephalopathies as neurodegenerative disorders. In: *Neurodegenerative Diseases: Pathology, Mechanisms, and Potential Therapeutic Targets*, pp. 295–315.
- Hu, T., Zhang, J., Wang, J., et al., 2023. Advances in epilepsy: mechanisms, clinical trials, and drug therapies. *J. Med. Chem.* 66, 4434–4467.
- Irwin, J.J., Shoichet, B.K., 2005. ZINC— a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* 45, 177–182.
- Josephson, C.B., Patten, S.B., Bulloch, A., et al., 2017. The impact of seizures on epilepsy outcomes: a national, community-based survey. *Epilepsia* 58, 764–771.
- Kim, S., Chen, J., Cheng, T., et al., 2019. PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res.* 47, D1102–D1109.
- Kondapuram, S.K., Sarvagalla, S., Coumar, M.S., 2021. Docking-based virtual screening using PyRx Tool: autophagy target Vps34 as a case study. In: *Molecular Docking for Computer-Aided Drug Design*, Elsevier, pp. 463–477.
- Kramer, O., Kramer, O., 2016. Scikit-learn. In: *Machine learning for evolution strategies*, pp. 45–53.
- Langeh, U., Singh, S., 2021. Targeting S100B protein as a surrogate biomarker and its role in various neurological disorders. *Curr. Neuropharmacol.* 19, 265–277.
- Liang, K.-G., Mu, R.-Z., Liu, Y., et al., 2019. Increased serum S100B levels in patients with epilepsy: a systematic review and meta-analysis study. *Front. Neurosci.* 13, 456.
- Liu, C.-H., Lin, Y.-W., Tang, N.-Y., et al., 2012. Neuroprotective effect of Uncaria rhynchophylla in kainic acid-induced epileptic seizures by modulating hippocampal mossy fiber sprouting, neuron survival, astrocyte proliferation, and S100b expression. *Evidence-Based Complement. Alternat. Med.*
- Lovrić, M., Molero, J.M., Kern, R., 2019. PySpark and RDKit: moving towards big data in cheminformatics. *Mol. Inf.* 38, 1800082.
- Macalino, S.J.Y., Gosu, V., Hong, S., et al., 2015. Role of computer-aided drug design in modern drug discovery. *Arch. Pharm. Res.* 38, 1686–1701.
- Macalino, S.J.Y., Basith, S., Clavio, N.A.B., et al., 2018. Evolution of *in silico* strategies for protein-protein interaction drug discovery. *Molecules* 23, 1963.
- Mysinger, M.M., Carchia, M., Irwin, J.J., et al., 2012. Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. *J. Med. Chem.* 55, 6582–6594.



- Noor, F., Noor, A., Ishaq, A.R., et al., 2021. Recent advances in diagnostic and therapeutic approaches for breast cancer: a comprehensive review. *Curr. Pharm. Des.* 27, 2344–2365.
- Noor, F., Tahir ul Qamar, M., Ashfaq, U.A., et al., 2022. Network pharmacology approach for medicinal plants: review and assessment. *Pharmaceuticals* 15, 572.
- Noor, F., Asif, M., Ashfaq, U.A., et al., 2023. Machine learning for synergistic network pharmacology: a comprehensive overview. *Brief. Bioinform.*, bbad120
- Patel, L., Shukla, T., Huang, X., et al., 2020. Machine learning methods in drug discovery. *Molecules* 25, 5277.
- Pitkänen, A., Lukasiuk, K., 2009. Molecular and cellular basis of epileptogenesis in symptomatic epilepsy. *Epilepsy Behav.* 14, 16–25.
- Prada Gori, D.N., Llanos, M.A., Bellera, C.L., et al., 2022. iRaPCA and SOMoC: development and validation of web applications for new approaches for the clustering of small molecules. *J. Chem. Inf. Model.* 62, 2987–2998.
- Richard, J.M., 2001. Rocking and rolling with Ca<sup>2+</sup> channels. *Trends Neurosci.* 24, 445–449.
- Rose, P.W., Prlić, A., Altunkaya, A., et al., 2016. The RCSB protein data bank: integrative view of protein, gene and 3D structural information. *Nucleic Acids Res.* gkw1000.
- Sadaqat, M., Qasim, M., ul Qamar, M.T., et al., 2023. Advanced network pharmacology study reveals multi-pathway and multi-gene regulatory molecular mechanism of *Bacopa monnieri* in liver cancer based on data mining, molecular modeling, and microarray data analysis. *Comput. Biol. Med.* 161, 107059.
- Samad, A., Ajmal, A., Mahmood, A., et al., 2023. Identification of novel inhibitors for SARS-CoV-2 as therapeutic options using machine learning-based virtual screening, molecular docking and MD simulation. *Front. Mol. Biosci.* 10, 1060076.
- Sandhu, H., Kumar, R.N., Garg, P., 2022. Machine learning-based modeling to predict inhibitors of acetylcholinesterase. *Mol. Divers.* 26, 331–340.
- Santos, B.S., Silva, I., da Câmara Ribeiro-Dantas, M., et al., 2020. COVID-19: A scholarly production dataset report for research analysis. *Data Brief* 32, 106178.
- Systèmes, D., 2019. Discovery Studio Visualizer. v16. 1, 15350.
- Tahir ul Qamar, M., Zhu, X.-T., Chen, L.-L., et al., 2022. Target-specific machine learning scoring function improved structure-based virtual screening performance for SARS-CoV-2 drugs development. *Int. J. Mol. Sci.* 23, 11003.
- Van Eldik, L.J., Griffin, W.S.T., 1994. S100 $\beta$  expression in Alzheimer's disease: relation to neuropathology in brain regions. *Biochimica et Biophysica Acta (BBA)-Mol. Cell Res.* 1223, 398–403.
- Wong, M., 2005. Modulation of dendritic spines in epilepsy: cellular mechanisms and functional implications. *Epilepsy Behav.* 7, 569–577.
- Yang, J., Cai, Y., Zhao, K., et al., 2022. Concepts and applications of chemical fingerprint for hit and lead screening. *Drug Discov. Today*, 103356.