Title

Screening of BindingDB database ligands against EGFR, HER2, Estrogen, Progesterone and NF- κ B receptors based on machine learning and molecular docking

Authors

Rezaee, Parham

Rezaee, Shahab

Maaza, Malik

et al.

Publication Date

2024-12-01

DOI

10.1016/j.compbiomed.2024.109279

Computers in Biology and Medicine 183 (2024) 109279

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Screening of BindingDB database ligands against EGFR, HER2, Estrogen, Progesterone and NF-κB receptors based on machine learning and molecular docking

Parham Rezaee [a,b], Shahab Rezaee [a], Malik Maaza [b], Seyed Shahriar Arab [c,*]

[a] Department of Biophysics, School of Biological Sciences, Tarbiat Modares University, Tehran, Iran
[b] UNESCO-UNISA-iTLABS Africa Chair in Nanoscience and Nanotechnology (U2ACN2), College of Graduate Studies, University of South Africa (UNISA), Pretoria, South Africa
[c] Department of Pediatrics, University of California, La Jolla, San Diego, 92093, CA, USA

ARTICLE INFO

Keywords:
Virtual screening
Machine learning
Molecular docking
Breast cancer

ABSTRACT

Breast cancer, the second most prevalent cancer among women worldwide, necessitates the exploration of novel therapeutic approaches. To target the four subgroups of breast cancer "hormone receptor-positive and HER2-negative, hormone receptor-positive and HER2-positive, hormone receptor-negative and HER2-positive, and hormone receptor-negative and HER2-negative" it is crucial to inhibit specific targets such as EGFR, HER2, ER, NF-κB, and PR.

In this study, we evaluated various methods for binary and multiclass classification. Among them, the GA-SVM-SVM:GA-SVM-SVM model was selected with an accuracy of 0.74, an F1-score of 0.73, and an AUC of 0.92 for virtual screening of ligands from the BindingDB database. This model successfully identified 4454, 803, 438, and 378 ligands with over 90% precision in both active/inactive and target prediction for the classes of EGFR+HER2, ER, NF-κB, and PR, respectively, from the BindingDB database. Based on to the selected ligands, we created a dendrogram that categorizes different ligands based on their targets. This dendrogram aims to facilitate the exploration of chemical space for various therapeutic targets.

Ligands that surpassed a 90% threshold in the product of activity probability and correct target selection probability were chosen for further investigation using molecular docking. The binding energy range for these ligands against their respective targets was calculated to be between −15 and −5 kcal/mol. Finally, based on general and common rules in medicinal chemistry, we selected 2, 3, 3, and 8 new ligands with high priority for further studies in the EGFR+HER2, ER, NF-κB, and PR classes, respectively.

1. Introduction

Breast cancer, characterized by the highest mortality rate among various cancer types, is a widespread condition [1]. Despite remarkable advancements in the fields of basic life sciences and biotechnology, the process of drug discovery and development (DDD) for breast cancer medicine remains slow and costly. On average, it takes around 15 years and approximately $2 billion to develop a small-molecule drug [2]. While clinical studies are widely acknowledged as the most expensive phase in drug development, the greatest potential for time and cost savings lies in the earlier stages of discovery and preclinical research. Preclinical efforts alone account for over 43% of pharmaceutical expenses, in addition to significant public funding [2–4]. This high cost is primarily due to the high attrition rate observed at every step, ranging from target selection and hit identification to lead optimization and the selection of clinical candidates. Furthermore, the substantial failure rate in clinical trials, currently at 90% [5], can largely be attributed to issues originating in the early stages of discovery, such as inadequate target validation or suboptimal properties of ligands. Identifying faster and more accessible methods to discover a broader range of high-quality chemical probes, hits, and leads with optimal absorption, distribution, metabolism, excretion, and toxicology (ADMET) as well as pharmacokinetics (PK) profiles during the early phases of DDD would significantly enhance outcomes in preclinical and clinical studies. Consequently, this would enable the development of more effective, accessible, and safer breast cancer drugs [5,6]. Understanding the specific receptors involved in breast cancer and the interactions of ligands that inhibit them is crucial for developing effective treatments.

* Corresponding author.
E-mail address: ssarab@health.ucsd.edu (S.S. Arab).

https://doi.org/10.1016/j.compbiomed.2024.109279
Received 26 June 2024; Received in revised form 24 September 2024; Accepted 14 October 2024
Available online 25 October 2024

Breast cancer is driven by the interaction of estrogen and progesterone receptors with breast cells [7]. These hormones, estrogen and progesterone, bind to their respective receptors, leading to dimerization and subsequent entry into the nucleus. Additionally, they bind to estrogen and progesterone response elements located near the promoters of target genes. Studies have shown that estradiol promotes the growth of breast cancer cells, while tamoxifen, an estrogen blocker, inhibits it. Targeting these hormone receptors can help identify potent inhibitors for hormone-mediated (ER+/PR+) breast cancer [8,9].

EGFR, a transmembrane glycoprotein, plays a crucial role in cell signaling and is a significant target in breast cancer treatment. EGFR activation leads to cell proliferation and differentiation [10], and anti-EGFR agents have shown efficacy in patients with specific mutations. Tyrosine kinase inhibitors (TKIs) such as gefitinib, erlotinib, and lapatinib are used to inhibit EGFR overexpression, providing therapeutic benefits in breast cancer [11,12]. Similarly, HER2, a protein with tyrosine kinase activity, is amplified in about 30% of human breast carcinomas and is linked to increased invasiveness and angiogenesis in breast cancer cells [13]. HER2 overexpression is linked to increased invasiveness and angiogenesis in breast cancer cells [14]. Targeting HER2 has led to the development of inhibitors like neratinib and afatinib, though further trials are necessary to confirm their efficacy. HER2's interaction with other receptors and its role in signaling pathways underscore its importance in breast cancer progression and treatment [15,16].

Nuclear factor-kappa B (NF-κB) is a transcription factor involved in cell proliferation, immune responses, and inflammation, contributing to the development of breast tumors [17]. In breast cancer, NF-κB activation occurs downstream of EGFR signaling, particularly in the ER-negative subtype. HER2 overexpression activates the PI3K/Akt pathway, leading to NF-κB induction and promoting angiogenesis through VEGF and IL-8 expression. It can activate two signaling pathways: the classical (canonical) pathway and the alternative (noncanonical) pathway [16,18]. Drugs like lapatinib and microtubule disruptors activate NF-κB, while studies have shown that ginseng inhibits COX-2 and NF-κB activation in breast cancer cell lines [19,20].

In the realm of computer-aided drug discovery, machine learning plays a pivotal role in elucidating the intricate relationships between chemical structures and biological activities, offering insights into optimizing compounds for enhanced binding affinity and biological responses [21,22]. This process involves utilizing chemical descriptors, which are numerical representations extracted from structures, and chemical fingerprints, high-dimensional vectors commonly employed in analysis and virtual screening applications. The construction of Quantitative Structure-Activity Relationship (QSAR) models follows a systematic protocol, encompassing steps such as molecular encoding, feature selection through unsupervised learning techniques, and the application of supervised machine learning models to establish mappings between input features and biological responses [23]. Evaluating the efficacy of QSAR models entails considerations like dataset selection, performance metrics assessment, and the incorporation of big data and machine learning to predict diverse biological phenomena. The evolution in this field underscores the necessity to move beyond traditional ligand-protein interaction-based drug design methods to meet contemporary clinical safety standards [24–26].

Molecular docking stands out as a prevalent and effective structure-based computational method used to forecast interactions between molecules and biological targets. This technique involves predicting the alignment of a ligand within a receptor and assessing their compatibility using a scoring mechanism [27]. Since its inception in the mid-1970s, docking has become indispensable for understanding molecular interactions, aiding in drug discovery, and facilitating development processes. Over time, there has been a notable surge in studies utilizing molecular docking to uncover essential structural elements for efficient ligand–receptor binding and to enhance the accuracy of docking methods. Noteworthy among these endeavors is a seminal study by Kuntz et al. in the early 1980s, reflecting the enduring significance and evolving sophistication of docking applications in drug discovery and biology [28]. The binding free energy is described as a sum of the intermolecular interactions between the ligand and the protein and the internal steric energy of the ligand. It can be represented by the equation:

$$\Delta G_{binding} = \Delta G_{vdw} + \Delta G_{H-bond} + \Delta G_{electrostatic} + \Delta G_{internal}$$

In this equation, $\Delta G_{binding}$ represents the total binding free energy, ΔG_{vdw} denotes the van der Waals interaction energy, ΔG_{H-bond} refers to the energy associated with hydrogen bonds, $\Delta G_{electrostatic}$ accounts for the electrostatic interactions, and $\Delta G_{internal}$ indicates the internal steric energy of the ligand. The van der Waals interaction is calculated using a LJ 6–12 potential between the protein and the ligand atoms. The steric part of the H-bond term is calculated using a LJ 10–12 potential. The intermolecular electrostatic interaction is calculated using Coulomb's law. The internal energy of the ligand is a sum of steric and electrostatic interactions calculated for non-bonded ligand atoms [29–32].

Recent advances in virtual screening for breast cancer drug discovery have highlighted several promising approaches. Awasthi et al. (2014) utilized a 3D-QSAR CoMFA model on flavonoids to target aromatase, a key enzyme in estrogen-dependent breast cancer, identifying 7- hydroxyflavanone beta-D-glucopyranoside as a potent inhibitor [33]. Yousuf et al. (2017) screened 3 million compounds against EGFR, HER2, and HSP90, discovering five compounds with strong binding energies and favorable ADMET properties [34]. Anbuselvam et al. (2020) focused on EGFR inhibitors, conducting virtual screening, ADME predictions, and molecular dynamics simulations to identify four promising compounds [35]. Tsou et al. (2020) demonstrated the superiority of deep neural networks (DNN) in identifying TNBC inhibitors, using DNN to find potent hits from a 165 000-compound database and expanding the method to GPCR agonists [36]. He et al. (2021) developed machine learning and deep learning models for breast cancer cell lines, achieving high predictive accuracy across various models and molecular representations [37]. Aziz et al. (2022) screened benzene sulphonamide derivatives, identifying compound 762 with superior binding affinity compared to Dabrafenib, supported by deep learning predictions [38]. Finally, Nada et al. (2023) developed a machine learning application to predict the bioactivity of EGFR inhibitors, leading to the synthesis of 18 novel compounds, with compound 9 showing significant antiproliferative and EGFR inhibitory activity, making it a promising candidate for breast cancer therapy [39].

In this study,we initiated by downloading all 3D structures from the BindingDB database, followed by optimizing these structures and generating 5668 descriptors for each molecule. Subsequently, we extracted ligands previously studied for their inhibitory effects on breast cancer targets (EGFR, ER, HER2, NF-κB, and PR) from this dataset. Through a series of data preprocessing procedures and the allocation of activity labels to each molecule, we developed two models to categorize molecules as active or inactive and to identify the target of each ligand. To construct these models, we utilized various feature selection algorithms to choose descriptors and employed diverse machine learning models, tuning their hyperparameters to identify the most effective models for virtual screening. After assessing and selecting the most effective models for these classifications, we established a pipeline using these models. This pipeline was then employed to screen the BindingDB ligands, utilizing different precision thresholds for the classifiers. The identified ligands underwent molecular docking to evaluate their binding energy with the respective targets. Additionally, we applied several established principles in medicinal chemistry to prioritize further investigation of these selected molecules, such as molecular dynamics, in vitro and in vivo studies. Furthermore, we examined the significance of the features employed in creating the target predictor model, aiming to identify a simple rule for acceptable accurate target recognition as a common rule. This study not only presented a robust model for identifying potential inhibitors for breast cancer targets but also introduced a

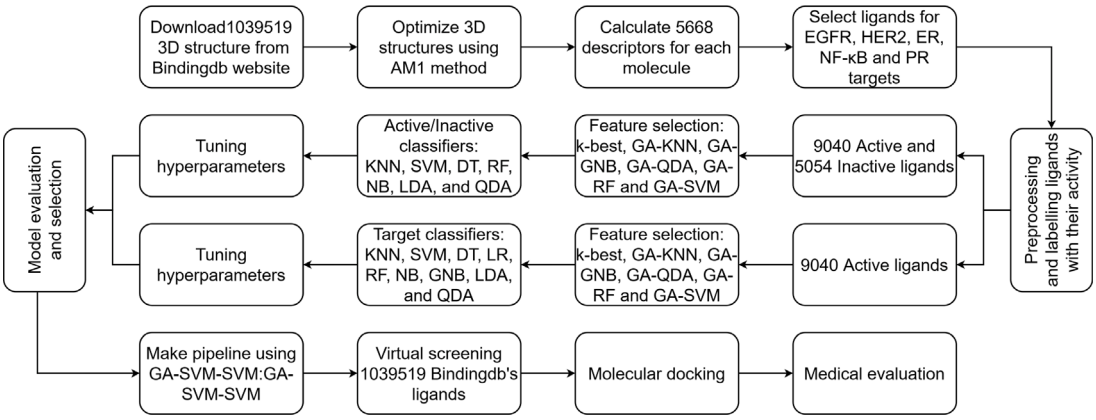

Fig. 1. Workflow diagram illustrating the sequential steps followed in the study. The process begins with the acquisition of 3D structures from the BindingDB database, followed by structure optimization and the generation of 5668 descriptors for each molecule. Subsequently, ligands with known inhibitory effects on EGFR, ER, HER2, NF-κB, and PR receptors, are extracted from this dataset. Through a series of data preprocessing steps and the assignment of activity labels, two models are developed to classify molecules as active/inactive and to determine the target of each ligand. Various feature selection algorithms are utilized to select descriptors, and diverse machine learning models are employed with hyperparameter tuning to identify the most effective models for virtual screening. Following the evaluation and selection of optimal models, a pipeline is constructed for screening the BindingDB ligands with different precision thresholds for the classifiers. Identified ligands then undergo molecular docking to assess their binding energy with respective targets. Additionally, established principles of medicinal chemistry are applied to prioritize further investigation of selected molecules.

dendrogram to assist researchers in navigating the chemical space. This dendrogram allows for the efficient identification of potential inhibitors by examining key molecular features, reducing the need to rerun the model while maintaining high accuracy. Furthermore, we proposed a list of ligands with potential repositioning properties to inhibit breast cancer targets, demonstrating their interactions with their respective targets. These ligands hold promise for further investigation, including in vitro and in vivo studies, to explore their potential in treating breast cancer.

Table 1
Number of active and inactive molecules for each class.

Target	Active	Inactive	Total
EGFR	4922	2419	**7341**
ER	1223	636	**1859**
HER2	1393	789	**2182**
NF-κB	447	826	**1273**
PR	1055	384	**1439**
Total	**9040**	**5054**	**14 094**

2. Materials and methods

As the procedure's workflow of this study is illustrated in Fig. 1, to obtain a dataset consisted of inhibitors targeting various breast cancer targets, we downloaded five sets specific sdf files (EGFR, HER2, ER, NF-κB, and PR inhibitors) from the Binding database website [40] (version 2022m8). The number of sdf files for each targets are: 7341 for EGFR, 2182 for HER2, 1859 for ER, 1273 for NF-κB, and 1439 for PR. Since near 70% of inhibitors in BindingDB database for EGFR and HER2 targets, were identical, we merged these two classes and named them EGFR + HER2 class. These sdf files were then converted to gjf files using the OpenBabel [41] software. The 3D structures of all the molecules were optimized using the Austin model 1 Hamiltonian implemented in Gaussian software [42], a program for electronic structure calculations. The optimized molecules were used to calculate molecular descriptors with the help of Alvadesc [43] software, which efficiently computes descriptors essential for QSAR/QSPR modeling and other cheminformatics applications. A total of 5668 descriptors (in type of float), including 0-, 1-, 2-, and 3D descriptors, were generated. To streamline the dataset in preprocessing step, descriptors with constant values in 90% of the compounds were removed. Additionally, among descriptors with a correlation above 0.9, the one exhibiting higher pair correlation with all other descriptors was kept and the others were automatically excluded. Following these processes, 1461 descriptors remained for further analysis.

Each sdf file contains activity information pertaining to a specific molecule, indicating the affinity of that molecule towards different therapeutic targets. We extracted the activity information from the downloaded sdf files for each class of molecules, and saved it in separate vectors. The enumeration of the collected data can be found in Table 1. Molecules with IC_{50}, K_i, and EC_{50} values below 2000 nM were categorized as active inhibitors, while those with values exceeding 2000 nM were considered inactive. Also we removed molecules with activity more than 10 000 nM as outlier data from the inactive dataset. The prepared dataset was used to construct active/inactive and target classifier. Both active and inactive molecules were utilized to develop and evaluate the active/inactive classifiers. These models serve the purpose of screening extensive databases and identifying new potent molecules for the treatment of breast cancer.

We employed various methods, including k-best, K-Nearest Neighbors (KNN), Gaussian Naive Bayes (GNB), Quadratic Discriminant Analysis (QDA), Random Forest (RF), and Support Vector Machine (SVM), to independently select suitable descriptors for the active/inactive and target classifiers. To optimize the selection of features across all data, we utilized a Genetic Algorithm (GA) in an optimal manner. The GA started with a population size of 200 and evolved through a maximum of 1000 generations, employing a crossover rate of 0.5 and a mutation rate of 0.2. The estimator was configured with the aforementioned methods, utilizing 5-fold cross-validation and an accuracy scoring function. The only difference between the feature selection processes of the active/inactive and target classifiers was the maximum number of features. The active/inactive classifier allowed a maximum of 128 features, while the target classifier allowed a maximum of 64 features.

We utilized the chosen features to create an optimized binary classifier for predicting active/inactive molecules. Various methods, KNN, SVM, decision tree (DT), RF, naive bayes (NB), linear discriminant analysis (LDA), and QDA, were employed for this purpose. There are several hyperparameters for each methods which should optimize to make ideal models. To maximize the performance of each method, we conducted a grid search to identify the best parameters for constructing the model. Table S1 demonstrated the different range of values of hyperparameters for each methods that are used in grid search algorithm. Given the balanced nature of our dataset and the limited availability of active molecule data, we selected 384 active and 384 inactive molecules randomly from each of the EGFR + HER2, ER, NF-κB, and PR classes without replacement. This resulted in a balanced

dataset consisting of 1536 active and 1536 inactive ligands, which was used to construct the binary classifier. Of this dataset, 70% was allocated for training the model, with the remaining 30% reserved for testing.

For the target classifier, we employed methods such as KNN, SVM, DT, logistic regression (LR), RF, NB, GNB, LDA, and QDA. Similar to the binary classifier, we utilized grid search to identify the optimal parameters for constructing the model (see Table S2 to find the range of values which are used to tune the model using grid search algorithm). To maintain a balanced dataset, we selected randomly 440 active molecules from each class (EGFR + HER2, ER, NF-κB, and PR) without replacement for use in the multi-class classifier. We then allocated 70% of this dataset for model training, with the remaining 30% designated for testing. In constructing multi-class classification models for ligand-based virtual screening, employing a variety of algorithms and tuning their hyperparameters using grid search allows for a thorough evaluation of model performance across different classification tasks.

Based on the results from each model, we selected the top two models for both active/inactive and target classification to create combined active/inactive:target models. These combined models were then evaluated using various combinations of the top two selected models for both classifications. Our objective was to identify the best model for virtual screening by assessing their performance across several key metrics: precision, recall, sensitivity, specificity, accuracy, F1-score, Matthews correlation coefficient (MCC), and area under the curve (AUC). Each combination was thoroughly tested to ensure it could effectively differentiate between active and inactive molecules and accurately classify the target receptors, providing a robust evaluation of their predictive capabilities for virtual screening.

A dataset of 1 039 519 molecules was gathered from the BindingDB database. The same preparation process used for the breast cancer inhibitors described earlier was applied to these downloaded molecules. Additionally, the descriptors selected for the breast cancer inhibitors were also chosen for these molecules. This resulted in a data matrix of size 1 039 519 × 1461, which was used for virtual screening. During virtual screening, first the active/inactive predictor assigned the active molecules in the class of "active" and the inactive ones placed in the class of "N/A". Then the target classifier assign the class of each molecules to EGFR + HER2, ER, NF-κB, or PR from the "active" class. The decision-making process for the model's predictions of activity and target was constrained by a certainty threshold of 0.0, 0.8, 0.85, and 0.9 for the active/inactive classifier, and 0.0, and 0.9 for the target classifier.

To evaluate the binding energy of selected molecules which are obtained from virtual screening, we employed molecular docking. This calculation helped us to verify that selected new ligands can bind to their therapeutic targets. Autodock Vina [44] was utilized for molecular docking to calculate the binding affinities between ligands and their respective targets. First, we converted the format of optimized structure of selected molecules to pdbqt using openbabel software to prepare ligands for molecular docking. Then we started the protein preparation stage by changing the format of proteins to pdbqt, too. The docking runs were conducted with an exhaustiveness parameter set to 32, ensuring thorough exploration of the conformational space, and 100 predicted poses were generated for each ligand-protein interaction. The grid box resolution was set with specific coordinates for each target: (EGFR)(PDB ID: 1M17) had coordinates of 23.424, 1.310, 51.002 along the x, y, and z axes, respectively, with a grid spacing of 0.2 Å; (HER2)(PDB ID: 3PP0) had coordinates of 17.563, 16.689, 26.321; (PR)(PDB ID: 1A28) had coordinates of 17.038, 0.145, 74.798; (ER)(PDB ID: 2IOK) had coordinates of 19.050, 35.696, 52.244; and (NF-κB)(PDB ID: 4KIK) had coordinates of 48.268, 31.589, −57.885. These coordinates were used to define the binding sites for the docking process. The grid dimensions were set at 25.2 × 25.2 × 25.2 Å. To efficiently manage the computational workload, each docking job was assigned to a single CPU core. The control ligands were initially docked with the binding sites of the five receptors, and the resulting interactions were compared with standard reference ligands.

Table 2
The evaluation of GA-RF-RF and GA-SVM-SVM for binary and GA-QDA-SVM and GA-SVM-SVM models for therapeutic classification.

Method	GA-RF-RF			
Best Param	bootstrap = True, criterion = 'gini', max_features = 32, n_estimators = 400, random_state = 42			
	Precision	Recall	f1-score	Support
Active	0.76	0.72	0.74	461
Inactive	0.74	0.77	0.75	461
Accuracy		0.75		922
Macro avg	0.75	0.75	0.75	922
Weighted avg	0.75	0.75	0.75	922
	Sensitivity	Specificity	MCC	AUC
	0.7744	0.7310	0.5059	0.8356
Method	GA-SVM-SVM			
Best Param	C = 10, gamma = 0.01, kernel = 'rbf', probability = True, random_state = 42			
	Precision	Recall	f1-score	Support
Active	0.74	0.76	0.75	461
Inactive	0.75	0.74	0.74	461
Accuracy		0.75		922
Macro avg	0.75	0.75	0.75	922
Weighted avg	0.75	0.75	0.75	922
	Sensitivity	Specificity	MCC	AUC
	0.7983	0.7007	0.5013	0.8214
Method	GA-QDA-SVM			
Best Param	C = 10, decision_function_shape = 'ovo', gamma = 0.01, kernel = 'rbf', probability = True, random_state = 42, max_iter = −1			
	Precision	Recall	f1-score	Support
EGFR+HER2	0.96	0.92	0.94	132
ER	0.93	0.95	0.94	132
NF-κB	0.93	0.95	0.94	132
PR	0.97	0.97	0.97	132
Accuracy		0.95		528
Macro avg	0.95	0.95	0.95	528
Weighted avg	0.95	0.95	0.95	528
	Sensitivity	Specificity	MCC	AUC
	0.9451	0.9451	0.9268	0.9835
Method	GA-SVM-SVM			
Best Param	C = 10, gamma = 0.01, kernel = 'rbf', decision_function_shape = 'ovo', probability = True, random_state = 42, max_iter = −1			
	Precision	Recall	f1-score	Support
EGFR+HER2	0.96	0.91	0.93	132
ER	0.92	0.95	0.93	132
NF-κB	0.94	0.97	0.96	132
PR	0.96	0.95	0.96	132
Accuracy		0.95		528
Macro avg	0.95	0.95	0.95	528
Weighted avg	0.95	0.95	0.95	528
	Sensitivity	Specificity	MCC	AUC
	0.9394	0.9401	0.9194	0.9830

Because undesirable pharmacokinetics and toxicity of candidate compounds are the main reasons for the failure of drug development, it has been widely recognized that ADMET should be evaluated as early as possible. To prioritize new ligands for further studies, such as molecular dynamics and others, we utilized various rules such as Lipinski, Pfizer, GSK, and golden triangle rules. Additionally, important parameters for drug production, including QED, SAscore, and MCE-18, were calculated using ADMETlab 2.0 [45]. Ligands that met all these

criteria were selected as high-priority candidates for further investigation. The interactions of these selected ligands with their target proteins were analyzed using Discovery Studio 2024.

The computational infrastructure employed in this study encompassed 10 virtual machines (VM) each equipped with a 24-core Xeon 2690 processor and 32 GB of RAM for the optimization of 3D structures. A laptop featuring an AMD Ryzen 7 4800H processor and 16 GB of RAM facilitated the computation of molecular descriptors, the development of machine learning models, and the implementation of virtual screening procedures. Furthermore, a server powered by a Ryzen 9 7950X processor and possessing 128 GB of RAM was utilized specifically for conducting molecular docking analyses.

3. Results and discussion

EGFR and HER2 receptors, shows 83.71% similarity in their residues using sequence alignment with BLOSUM weight matrix and have a large similarity in their 3D structure using Needleman-Wunsch alignment algorithm with BLOSUM-62 similarity matrix (Figure S1 and S2). Moreover, near 70% of ligands in BindingDB database with EGFR and HER2 targets, were identical. According to these reasons, we merged two classes of EGFR and HER2 to just one EGFR + HER2 class.

To select the best features for active/inactive and target classifiers, we hired k-best, GA-KNN, GA-GNB, GA-QDA, GA-RF, and GA-SVM methods. The active/inactive and target classifiers utilized 128 and 64 selected features, respectively (see Tables S3 and S4). Each feature selection method offers unique advantages and disadvantages. K-best is computationally efficient but may overlook feature interdependencies. KNN is versatile but computationally expensive and sensitive to scaling. GNB is fast and handles high-dimensional data but assumes feature independence. QDA offers flexible decision boundaries but can overfit and is sensitive to normality assumptions. RF is robust to overfitting but computationally intensive. SVM is effective in high-dimensional spaces but requires careful parameter tuning. Integrating GA with these methods optimizes feature subsets by considering feature interactions, though it is computationally demanding.

According to each set of 128 selected features, KNN, SVM, DT, RF, NB, LDA, and QDA methods, created binary classifiers to recognize active and inactive molecules. The average values for sensitivity, specificity, accuracy, F1-score, MCC, and AUC across all 42 models were 0.68, 0.66, 0.67, 0.67, 0.34, and 0.73, respectively. Also, each set of 64 selected features are used to create different target classifiers using KNN, SVM, DT, LR, RF, NB, GNB, LDA, and QDA methods. The average values for these target classifiers across 60 models were 0.86, 0.87, 0.86, 0.86, 0.81, and 0.96, respectively. The result of optimized models (using grid search) with each set of selected features are comprehensively demonstrated in Table S5–S16. KNN is simple and interpretable but computationally intensive and sensitive to scaling. SVM (either binary or multi-class strategies) handles high-dimensional data well but is computationally expensive and requires careful tuning. DTs are intuitive but prone to overfitting. RF improves generalization but is computationally heavy. NB and GNB are fast but can perform suboptimally with correlated features. LR handles multi-class classification well via methods like softmax but is limited by linear decision boundaries. LDA works well with linearly separable data but struggles with complex relationships. QDA provides flexible decision boundaries but can overfit and requires more data. Grid search ensures optimized model performance but is computationally demanding and time-consuming. After thorough comparison across various metrics, sensitivity, specificity, accuracy, F1-score, MCC, and AUC, the top two models identified for the active/inactive classification were GA-SVM-SVM and GA-RF-RF, and for the target classification, GA-SVM-SVM and GA-QDA-SVM, as detailed in Tables S17–S18.

The GA-SVM-SVM binary classifier was constructed using the radial basis function (RBF) kernel with a gamma value of 0.1 and a regularization term of 1. This configuration ensures a balance between the smoothness of the decision boundary and the correct classification of training examples. Similarly, the GA-RF-RF binary classifier was built using the Gini impurity function with 400 trees and a maximum depth until all leaves were pure. Although 128 features were initially selected for the random forest model, it ultimately utilized only 32 features. The construction of this random forest involved bootstrap sampling, which helps in reducing variance and improving model robustness (see Table S5–S10 to see the evaluation of optimal models with each selected features for binary classification).

For the target classifier, GA-SVM-SVM and GA-QDA-SVM multi-class classifiers were employed. Both classifiers utilized the RBF kernel, a gamma value of 0.01, and a regularization term of 10. The classification was performed using the one-vs-one strategy, which compares each pair of classes separately. This strategy is known to provide higher prediction accuracy compared to the one-vs-rest approach, as it considers the decision boundaries between each pair of classes individually [46,47](see Table S11–S16 to see the evaluation of optimal models with each selected features for multi-class classification).

Table 3
The evaluation of GA-SVM-SVM:GA-SVM-SVM, GA-RF-RF:GA-QDA-SVM, GA-SVM-SVM:GA-QDA-SVM, and GA-RF-RF:GA-SVM-SVM models in the pipeline.

GA-SVM-SVM:GA-SVM-SVM	Precision	Recall	f1-score	Support
N/A	0.75	0.74	0.74	461
EGFR+HER2	0.68	0.74	0.71	99
ER	0.83	0.83	0.83	115
NF-κB	0.61	0.61	0.61	123
PR	0.78	0.79	0.78	124
Accuracy		0.74		922
Macro avg	0.73	0.74	0.73	922
Weighted avg	0.74	0.74	0.74	922
MCC			0.62	
AUC			0.92	
GA-RF-RF:GA-QDA-SVM	Precision	Recall	f1-score	Support
N/A	0.73	0.76	0.74	461
EGFR+HER2	0.68	0.70	0.69	99
ER	0.82	0.83	0.83	115
NF-κB	0.56	0.50	0.53	123
PR	0.80	0.76	0.78	124
Accuracy		0.73		922
Macro avg	0.72	0.71	0.71	922
Weighted avg	0.72	0.73	0.73	922
MCC			0.61	
AUC			0.93	
GA-SVM-SVM:GA-QDA-SVM	Precision	Recall	f1-score	Support
N/A	0.75	0.74	0.74	461
EGFR+HER2	0.68	0.73	0.70	99
ER	0.82	0.84	0.83	115
NF-κB	0.60	0.59	0.60	123
PR	0.78	0.78	0.78	124
Accuracy		0.74		922
Macro avg	0.73	0.74	0.73	922
Weighted avg	0.74	0.74	0.74	922
MCC			0.62	
AUC			0.92	
GA-RF-RF:GA-SVM-SVM	Precision	Recall	f1-score	Support
N/A	0.73	0.76	0.74	461
EGFR+HER2	0.69	0.71	0.70	99
ER	0.82	0.82	0.82	115
NF-κB	0.57	0.51	0.54	123
PR	0.80	0.77	0.78	124
Accuracy		0.73		922
Macro avg	0.72	0.71	0.72	922
Weighted avg	0.73	0.73	0.73	922
MCC			0.61	
AUC			0.93	

Table 4
The reported classification models for breast cancer targets.

Year	Targets	Data set		Method	Descriptors	Model validation	Statistical results	Refs.
		Train	Test					
2024	EGFR, HER2, ER, NF-κB, and PR	2150	922	GA-SVM-SVM: GA-SVM-SVM	Alvadesc	5-fold CV	Acc = 0.74, MCC = 0.62, AUC = 0.92	This work
2024	VEGFR-2	518	143	Adaboost	Discovery Studio 2020	10-fold CV	Acc = 0.83	[48]
2024	TBK1	959	303	Extra Tree Classifier	Mordred	12-fold CV	Acc = 0.88, AUC = 0.89	[49]
2024	ErbB1	5215	1738	XGBoost	Dragon descriptors	NA	Acc = 0.85, AUC = 0.92	[50]
2019	BCRP	2240	559	SVM	MOE descriptors and Pubchem fingerprints	5-fold CV	Acc = 0.96, MCC = 0.81, AUC = 0.96	[51]
2016	Estrogen receptor beta	474	237	MACCSFP-RF	PaDEL Descriptor	5-fold CV	Acc = 0.88, MCC = 0.75, AUC = 0.95	[52]
2015	BCRP	197	99	ANN	Dragon descriptors	NA	Acc = 0.74, MCC = 0.46	[53]

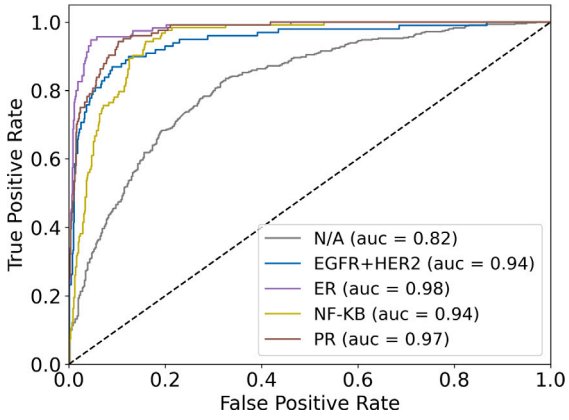

Fig. 2. The ROC plots for different classes with one-vs-rest strategy.

As shown in Table 2, the active/inactive classifiers GA-SVM-SVM and GA-RF-RF achieved precision, recall, and F1-scores all above 0.7, indicating a reliable performance in distinguishing between active and inactive molecules. Moreover, the high values in sensitivity, specificity, MCC and AUC validate the performance of classifiers. Similarly, the target classifiers GA-SVM-SVM and GA-QDA-SVM achieved precision, recall, and F1-scores all above 0.9, demonstrating high accuracy in classifying the ligands according to their target receptors. This high level of performance underscores the effectiveness of these optimized classifiers in ligand-based virtual screening. Moreover, the values upper than 0.9 for all sensitivity, specificity, MCC and AUC metrics validate the effectiveness of classifiers.

Subsequently, we generated a pipeline by combining different permutations of the selected model. The precision, recall, f1-score, and support of these models are presented in Table 3. Among the options, the GA-SVM-SVM:GA-SVM-SVM model emerged as the most suitable pipeline, displaying superior performance compared to others. This model could recognize the inactive molecules with the precision, recall, and f1-score more than 0.74. Moreover, this model could categorize the active ones with high f1-score 0.83 and 0.78 for the classes of ER and PR, respectively. Also, f1-score for two classes of EGFR + HER2 and NF-κB are quite good enough for accurate virtual screening. In overall, this approach achieved an accuracy of 0.74 and an AUC of 0.92, indicating its robust predictive capabilities. Moreover, this model achieved an MCC of 0.62, reflecting a robust performance that balances sensitivity and specificity. Fig. 2 showcases the receiver operating characteristic (ROC) plots for each class using the one-vs-rest strategy, further validating the effectiveness of the GA-SVM-SVM:GA-SVM-SVM model for virtual screening. It should be mentioned that the training an SVM with the RBF kernel is $O(n^2 \times m)$, where n is the number of training samples and m is the number of features. The SVM method with an RBF kernel offers several advantages over other methods. It provides high accuracy on challenging datasets with non-linear decision boundaries, exhibits good generalization performance, and effectively classifies unseen data better than classifiers like neural networks (NN) and DTs, due to its flexibility, robustness, and scalability. Additionally, SVMs can handle high-dimensional spaces well. However, SVMs also have some disadvantages. Unlike simpler models like NB or LR, SVMs require careful tuning of parameter values for optimal performance and are computationally expensive compared to similar models.

Table 5
Number of selected BindingDB molecules for each targets according to the threshold of 0, 80, 85, and 90% decision certainty for active/inactive prediction and the threshold of 0 and 90% decision certainty for target prediction.

Classes	Threshold				
	0:0	0:90	80:90	85:90	90:90
EGFR+HER2	172 498	95 123	19 796	11 068	4454
ER	54 101	22 876	3613	2029	803
NF-κB	45 452	16 400	2499	1257	438
PR	67 323	14 109	2300	1116	378

In recent years, numerous studies have leveraged machine learning methods to identify potential therapeutic compounds targeting various proteins implicated in breast cancer (see Table 4). For instance, Ding et al. [48] (2024) utilized an Adaboost model to screen ligands against VEGFR-2, achieving an accuracy of 0.83. Similarly, Siddiqui et al. [49] (2024) implemented an Extra Tree Classifier for TBK1, reporting an accuracy of 0.88 and an AUC of 0.89. Bouchama et al. [50] (2024) employed XGBoost to target ErbB1, with results showing an accuracy of 0.85 and an AUC of 0.92. Earlier studies also demonstrated success with different targets and methods; for example, Jiang et al. [51] (2019) used an SVM model for BCRP, achieving a high accuracy of 0.96 and an MCC of 0.81, while Niu et al. [52] (2016) applied MACCSFP-RF to predict ligands for Estrogen Receptor beta, obtaining an accuracy

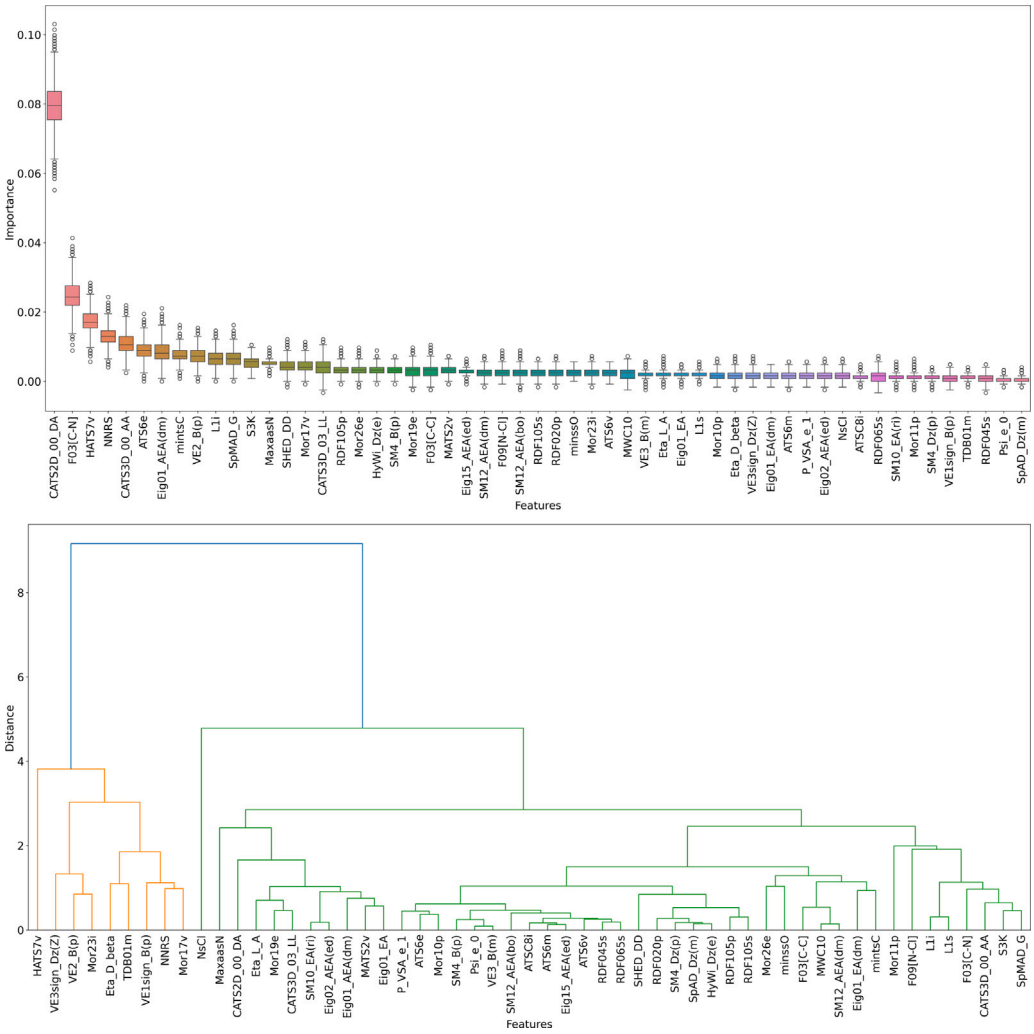

Fig. 3. The left plot shows the importance of features using permutation importance method using GA-SVM-SVM model for target prediction and the right one demonstrates hierarchical clustering dendrogram using pearson method to find the correlation distance of each features.

of 0.88 and an AUC of 0.95. These studies collectively underscore the potential of machine learning-based virtual screening in accelerating drug discovery for breast cancer, highlighting the effectiveness of various algorithms and descriptors across different contexts. However, it is important to note that these prior works focused on single-target models. In contrast, this study not only provides a model with exceptional performance but also offers the capability to classify inhibitors across multiple targets.

After the development of classifiers and selecting GA-SVM-SVM:GA-SVM-SVM for the pipeline in order to find the active molecules at the first stage and after that categorized them to four classes of EGFR + HER2, ER, NF-κB, and PR classes, we ran the virtual screening for BindingDB dataset including 039 519 ligands with certainty threshold of 0.0, 0.8, 0.85, and 0.9 for the active/inactive classifier, and 0.0, and 0.9 for the target classifier. Table 5 provides insights into the number of selected molecules from the BindingDB database for each target, based on different predetermined thresholds. As it can be seen in this table, 339 374 molecules are categorized in these four classes in the absence of any thresholds for certainty of decision making. This pipeline can work with more accuracy employing higher thresholds for certainty of decision making. Notably, this table reveals the presence of 4454, 803, 438, and 378 new inhibitor molecules for EGFR + HER2, ER, NF-κB, and PR, respectively. These novel inhibitors were selected with 90% precision in both the active/inactive and therapeutic classification decision-making processes. These amount of ligands are just 0.58% of whole molecules in BindingDB database. This high level of precision ensures that the selected molecules are highly likely to be true inhibitors, making them valuable candidates for further validation and drug development efforts. The ability of the GA-SVM-SVM:GA-SVM-SVM model to accurately classify and select potential inhibitors highlights its efficacy in ligand-based virtual screening and its potential impact on accelerating the drug discovery process. It should be mentioned that the virtual screening with this model is $O(1039519 \times m)$, where 1039519 is number of molecules and m is the number of features.

In order to easily identify the target for each molecule, our objective was to extract a straightforward rule from the model. To achieve this, we employed permutation importance to determine the significance of each feature in the model. Additionally, we utilized the Pearson method to create a hierarchical clustering dendrogram, which helped us identify the correlation distance (Euclidean distance) between features. Fig. 3 displays the feature importance and hierarchical clustering dendrogram. Based on these findings, we constructed a simple questionnaire dendrogram for determining the target of each molecule which are selected with 90% precision in both the active/inactive and target classification decision-making, as illustrated in Fig. 4. The data presented in Fig. 4 provide concise and effective structure–activity relationship (SAR) information regarding the inhibitors. For instance, NF-κB inhibitors exhibit significantly lower values for molecular walk count of order 10 (MWC10) and signal 10/weighted by polarizability (Mor10p) compared to other inhibitors. Additionally, EGFR and

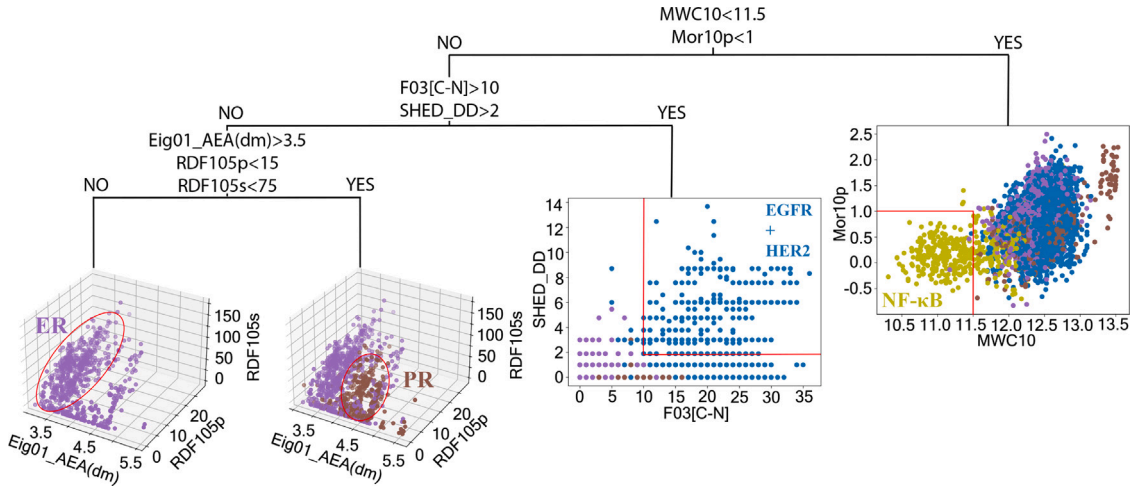

Fig. 4. A simple questionnaire dendrogram to separate ligands with a number of features and determine the targets of them.

HER2 inhibitors demonstrate higher values of frequency of C–N at topological distance 3 (F03[C-N]) and SHED Donor-Donor (SHED_DD) in comparison to ER and PR inhibitors. The ratio of eigenvalue n.1 from augmented edge adjacency mat. weighted by dipole moment (Eig01_AEA(dm)), radial distribution function–105/weighted by polarizability (RDF105p), and radial distribution function–105/weighted by I-state (RDF105s) differentiates PR and ER inhibitors. These SAR information types effectively filter a significant portion of large databases, thus accelerating early-stage drug discovery projects that begin with extensive databases like GDB-13 [54]. The classification of molecules based on their therapeutic targets has garnered considerable attention in the field of chemoinformatics [55]. These types of classifiers expand on the concept of "Chemography"[55,56], which refers to the art of navigating through chemical space. As evident from these figures, the inhibitors cluster within specific regions of the selected chemical space, aligning with the core objective of chemography.

In order to assess the binding energy of the molecules selected using the GA-SVM-SVM:GA-SVM-SVM model, we employed molecular docking for both the chosen molecules (with a multiplication of precision product exceeding 0.9 for both active/inactive and target classification) and the molecule sets within each class. The distribution of binding energy for these molecules, based on their molecular weights, is depicted in Fig. 5. In these plots, the pale dots represent the active inhibitors labeled by the bindingDB database, while the filled dots represent the active molecules utilized in constructing the GA-SVM-SVM:GA-SVM-SVM model. The red dots correspond to new inhibitors, which exhibit binding energy within the range of −15 to −5 kcal/mol. This range of binding energy proves to be sufficiently suitable for forming protein-ligand complexes. The average value of binding energy for each target of EGFR, HER2, ER, NF-κB, and PR for new ligands which are founded by the method are −9.48, −9.37, −8.97, −8.01, and −7.63 kcal/mol, respectively.

In order to prioritize further study on the new molecules, we applied several medicinal criteria. The Lipinski rule suggests that ligands with a molecular weight of less than or equal to 500 ($M_W \leq 500$), a logarithm of the n-octanol/water distribution coefficient of less than or equal to 5 ($logP \leq 5$), a number of hydrogen bond acceptors of less than or equal to 10 ($H_{acc} \leq 10$), and a number of hydrogen bond donors of less than or equal to 5 ($H_{don} \leq 5$) exhibit good absorption or permeability. Accordingly, 376, 59, 91, and 35 ligands were accepted based on the Lipinski rule for EGFR + HER2, ER, NF-κB, and PR targets, respectively. The Pfizer rule states that ligands with a $logP \geq 3$ and a topological polar surface area of less than or equal to 75 ($TPSA \leq 75$) are likely to be toxic. Consequently, 577, 30, 38, and 15 ligands passed the Pfizer rule for EGFR + HER2, ER, NF-κB, and PR targets, respectively. The GSK rule suggests that ligands with a $MW \leq 400$ and $logP \leq 4$ may have a more favorable ADMET (absorption, distribution, metabolism, excretion, and toxicity) profile. Thus, 6, 11, 65, and 12 ligands were accepted based on the GSK rule for EGFR + HER2, ER, NF-κB, and PR targets, respectively. Additionally, the golden triangle hypothesis proposes that ligands with a $200 \leq MW \leq 500$ and a logD (logarithm of the n-octanol/water distribution coefficient at pH = 7.4) ranging from −2 to 5 ($-2 \leq logD \leq 5$) may have a more favorable ADMET profile. Consequently, 166, 53, 91, and 34 ligands fulfilled the golden triangle criteria for EGFR + HER2, ER, NF-κB, and PR targets, respectively. Moreover, several parameters such as QED (desirability functions based on eight drug-likeness related properties including MW, $logP$, N_{HBA}, N_{HBD}, PSA, N_{rotb}, N_{Ar}), SAscore (synthetic accessibility score based on a combination of fragment contributions and a complexity penalty), and MCE-18 (medicinal chemistry evolution in 2018 score molecules by novelty in terms of their cumulative sp3 complexity) were considered favorable in the medical industry. Ligands with QED greater than 0.67, SAscore less than 6, and MCE-18 larger than 45 were deemed desirable. Accordingly, 6, 14, 4, and 14 ligands met these criteria for EGFR + HER2, ER, NF-κB, and PR targets, respectively. The distributions of selected molecules according to these rules are illustrated in Figures S3–S7. The molecules depicted in Fig. 6 satisfy all these criteria. Specifically, 2 (BDBM188812 and BDBM216448), 3 (BDBM149740, BDBM50189075, and BDBM50487909), 3 (BDBM483950, BDBM50221641, and BDBM 50460119), and 8 (BDBM400173, BDBM400179, BDBM400207, BDB M400217, BDBM481622, BDBM513587, BDBM513605, and BDBM 513606) ligands met all the parameters for EGFR + HER2, ER, NF-κB, and PR targets, respectively. As observed in this figure, each ligand within each class exhibits unique structural properties.

The top hits exhibited high binding energies with their targets, as shown in Table 6. Post-docking analysis revealed that all compounds effectively bound within the target domains, surrounded by key interacting residues and demonstrating notably high binding energy values (Figure S8). Figure S8 illustrates the molecular surface of the target binding pockets with their respective ligands in stick format. Additionally, 2D interaction plots in Figure S9 highlight significant binding-site interactions between the ligands and the targets' binding-site residues. These interactions are also listed in Table 6. As shown in the table, most interactions fall into categories such as hydrogen bonds, electrostatic, hydrophobic, halogen, and miscellaneous, all of which are favorable for enhancing the binding energy of the protein-ligand complexes. However, BDBM50460119 has an acceptor/donor clash with CYS99 in NF-κB, which is unfavorable for binding energy. Additionally, BDBM400217, BDBM513605, and BDBM513606 exhibit charge repulsion with ARG766 in the PR receptor.

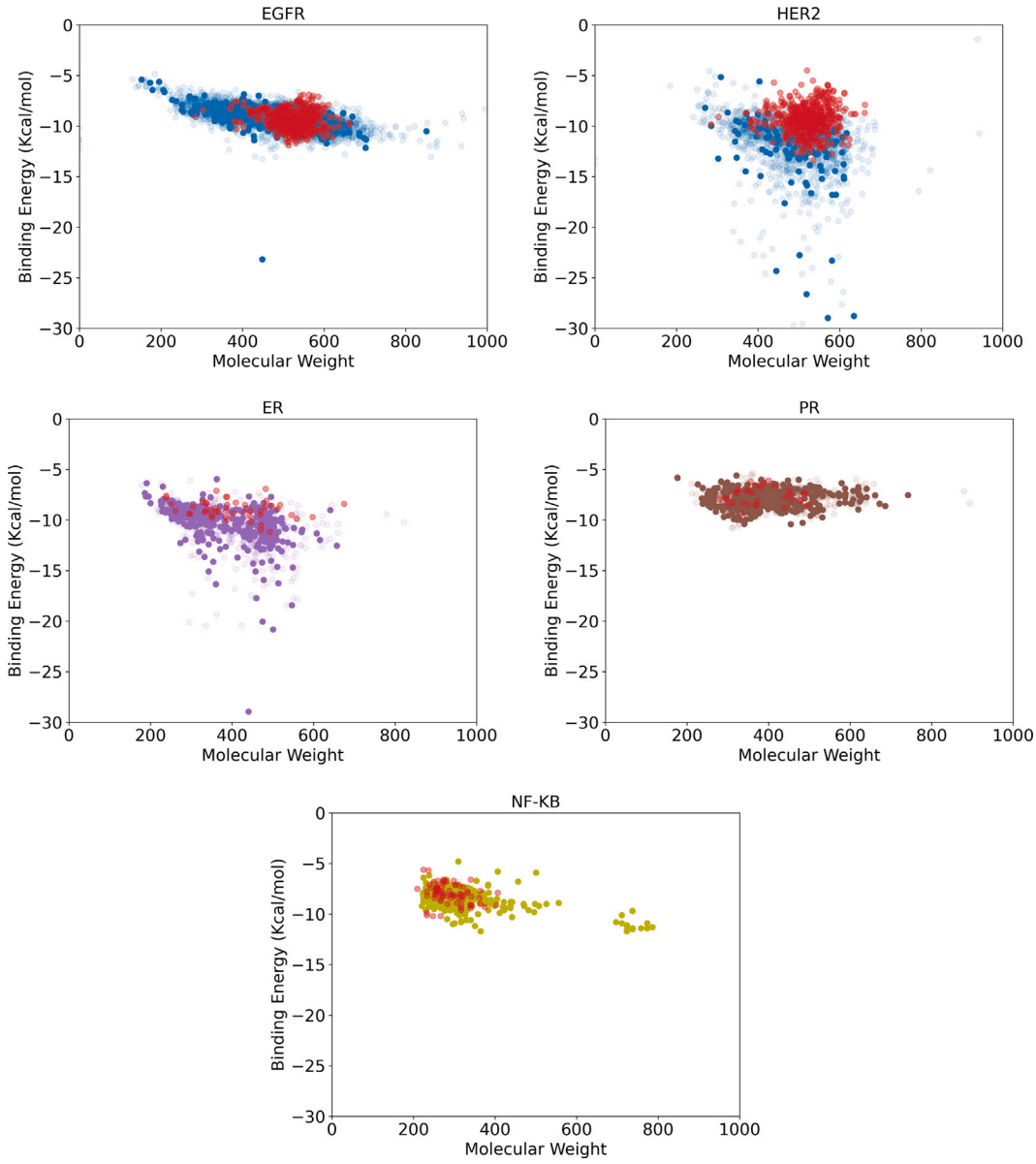

Fig. 5. Docking results of new ligands obtained from virtual screening. The pale dots in the following plots represent the active molecules in the BindingDB database for each class, the filled dots represent the molecules that participated in the construction of the model, and the red dots are the new molecules proposed by the model obtained from the screening.

4. Conclusion

In this study, we curated five sets of specific SDF files from the BindingDB database, targeting inhibitors of EGFR, HER2, ER, NF-κB, and PR. The datasets were randomly undersampled based on their activity and target classes. After filtering out constant and correlated descriptors, we selected 128 descriptors for active/inactive classification and 64 descriptors for target classification using k-best, GA-KNN, GA-GNB, GA-QDA, GA-RF, and GA-SVM methods. We then constructed binary classifiers, KNN, SVM, DT, RF, NB, LDA, and QDA, using the selected features to distinguish active from inactive molecules. Similarly, target classifiers were developed using KNN, SVM, DT, LR, RF, NB, GNB, LDA, and QDA methods, with all models optimized via grid search. The GA-SVM-SVM and GA-RF-RF models were identified as the top performers for active/inactive classification, while GA-SVM-SVM and GA-QDA-SVM excelled in target classification. These models achieved precision, recall, F1-scores, and AUC values above 0.7 and 0.9, respectively, underscoring their efficacy in ligand-based virtual screening.

A pipeline was then constructed by combining different permutations of the selected models. The GA-SVM-SVM:GA-SVM-SVM pipeline emerged as the most effective, achieving precision, recall, and F1-scores above 0.74 for inactive molecules. Furthermore, this model accurately categorized active molecules, with F1-scores of 0.83 and 0.78 for ER and PR classes, respectively. F1-scores for EGFR + HER2 and NF-κB were also sufficiently high for reliable virtual screening. The accuracy of this approach was 0.74, with an AUC of 0.92, demonstrating robust predictive capabilities. Using this model, we screened the BindingDB database and identified 4454, 803, 438, and 378 new inhibitor molecules for EGFR + HER2, ER, NF-κB, and PR targets, respectively, achieving 90% precision in both active/inactive and target classifications. The computational complexity of this algorithm was determined to be $O(n^2 \times m + 1039519 \times m)$. Additionally, we provided a simple dendrogram to aid in determining the target of new ligands, offering valuable insights into the distribution of these molecules within chemical space.

Molecules with precision scores exceeding 0.9 for both classifications were subjected to molecular docking analysis, revealing binding

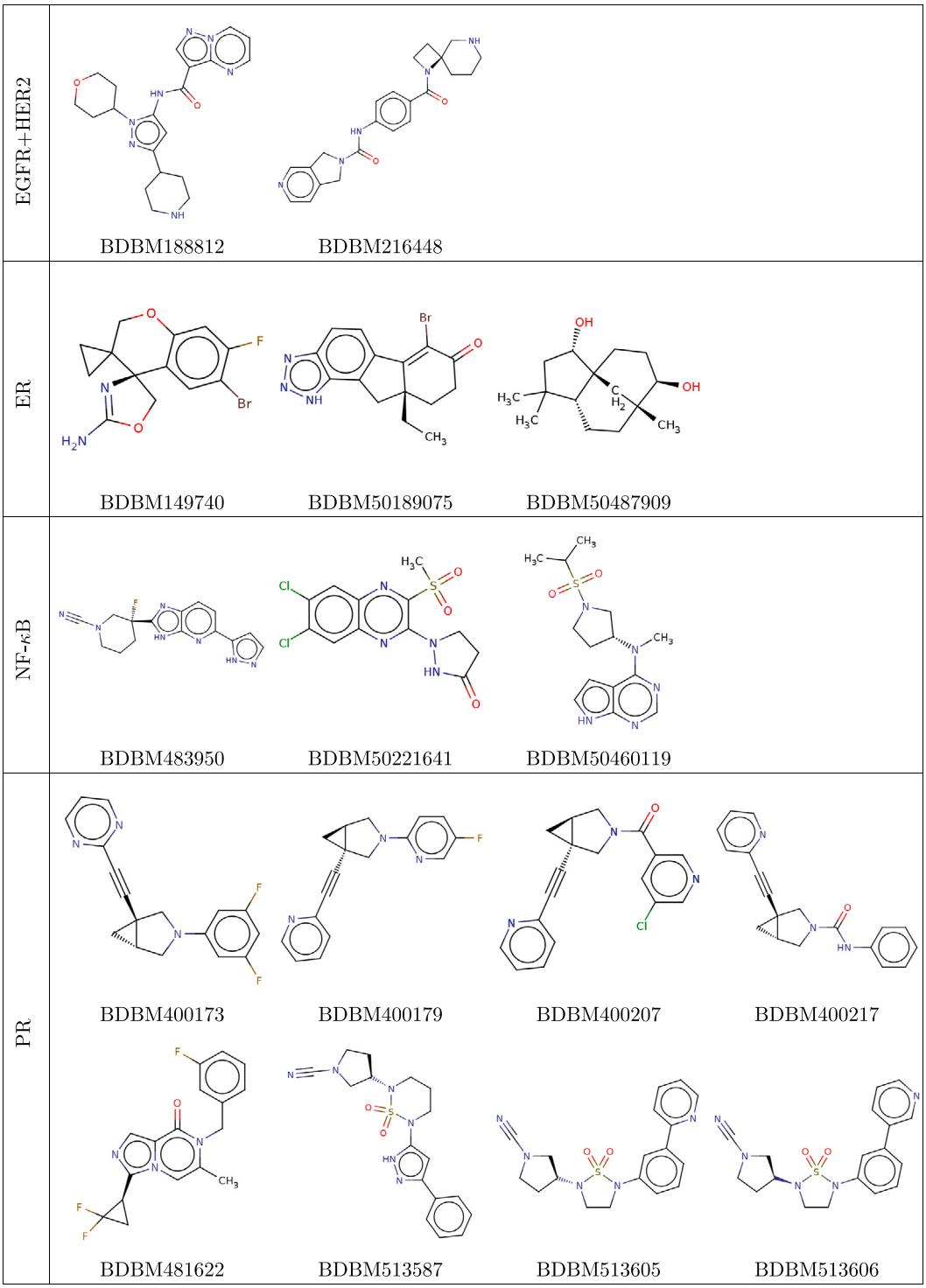

Fig. 6. List of the molecules which are met all criteria.

energies ranging from −15 to −5 kcal/mol, indicative of strong inhibitory potential. Further prioritization using Lipinski's rule, Pfizer, GSK, and golden triangle rules, along with QED, SAscore, and MCE-18 parameters, identified 2 (BDBM188812 and BDBM216448), 3 (BDBM149740, BDBM50189075, and BDBM50487909), 3 (BDBM483950, BDBM50221641, and BDBM50460119), and 8 (BDBM400173, BDBM400179, BDBM400207, BDB M400217, BDBM481622, BDBM513587, BDBM513605, and BDBM513606) ligands that met all criteria for EGFR+ HER2, ER, NF-κB, and PR targets, respectively. This research not only provides a comprehensive benchmark of machine learning methods for computational drug discovery but also introduces a set of novel ligands with promising inhibitory effects on breast cancer targets. Finally, the interactions of top hits with their respective targets were analyzed in detail.

Future work will aim to reproduce these results using more advanced machine learning techniques, such as deep learning and transfer learning, to further enhance the accuracy of our virtual screening pipeline. Validation of the binding energies and interactions of the

Table 6
Binding energy and molecular interaction of selected molecules with their targets.

Target	Inhibitor	Binding energy (kcal/mol)	Favorable					Unfavorable	
			Hydrogen bonds	Electrostatic	Hydrophobic	Halogen	Miscellaneous	Charge repulsion	Acceptor/ donor clash
EGFR	BDBM188812	−9.6	MET769, ASP831, GLU738, ASN818	ASP831	VAL702, PHE699, LEU820, LYS721, LEU694, ALA719, ARG817				
EGFR	BDBM216448	−9.6	LEU694, THR766	LYS721, ASP831	LEU694, VAL702, LEU764				
HER2	BDBM188812	−8.1	SER728, MET801, ARG849		LEU726, VAL734, ALA751, LEU852, LYS753, CYS805				
HER2	BDBM216448	−8.2	LEU726, ASP808, THR862, ASP863		VAL734, ALA751, LYS753, LEU796, CYS805				
ER	BDBM149740	−8.2			LEU349, ALA350, LEU384, LEU387, MET388, LEU391, PHE404, ILE424	LEU387			
ER	BDBM50189075	−9.2			LEU346, ALA350, MET421, ILE424		MET343		
ER	BDBM50487909	−7.6	GLY521		LEU346, LEU525				
NF-κB	BDBM483950	−8.4	CYS99	ASP166	LEU21, VAL29, ALA42, LYS44, CYS99, VAL152, ILE165				
NF-κB	BDBM50221641	−8.4			LEU21, VAL29, ALA42, TYR98, CYS99, VAL152, ILE165				
NF-κB	BDBM50460119	−7.5	THR23, GLU97		LEU21, ALA42, VAL74, CYS99, VAL152, ILE165				CYS99
PR	BDBM400173	−7.9		ARG766	PRO696, TRP765, ARG766, PHE818	LEU758			
PR	BDBM400179	−8	GLU695, LEU758, LYS769	LYS822	PRO696, VAL729, TRP732, LEU758, TRP765, ARG766				
PR	BDBM400207	−7.3	PRO696, ASP697	GLU695, ARG766	MET692, PRO696				
PR	BDBM400217	−8.5	PRO696	GLU695, LYS822	VAL729, TRP732, LEU758, TRP765, ARG766, LYS769			ARG766	
PR	BDBM481622	−9.3	VAL698, GLY762	GLU695, ARG766, LYS822	PRO696, VAL698, VAL729, TRP732, LEU758, ARG766	PRO696			
PR	BDBM513587	−8.1	GLY762	GLU695, ARG766	ARG766		TRP765		
PR	BDBM513605	−8.6	PRO696, ARG766	GLU695, ARG766, LYS822	PRO696, VAL729, TRP732, LEU758			ARG766	
PR	BDBM513606	−8.5	PRO696, ARG766, GLN815	GLU695, ARG766, LYS822	PRO696, VAL729, TRP732, LEU758			ARG766	

identified ligands through molecular dynamics simulations, as well as experimental analyses like in vitro and in vivo studies, will be crucial. We encourage other researchers to experimentally validate our models, as this study offers significant insights for the discovery of new inhibitors for breast cancer.

CRediT authorship contribution statement

Parham Rezaee: Writing – original draft, Visualization, Validation, Methodology, Formal analysis, Data curation, Conceptualization. **Shahab Rezaee:** Validation, Methodology, Formal analysis, Data curation, Conceptualization. **Malik Maaza:** Writing – review & editing, Resources, Conceptualization. **Seyed Shahriar Arab:** Writing – review & editing, Resources, Supervision, Conceptualization.

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.

Acknowledgments

The authors acknowledge the UNESCO UNISA iThemba-LABS/NRF Africa Chair in Nanoscience & Nanotechnology (U2ACN2) and the Centre for High-Performance Computing (CHPC), South Africa for providing computational resources and facilities for this research project. The authors would like to express their gratitude to all the members of the bioinformatics lab at Tarbiat Modares University (TMU) for their

valuable contributions in the form of discussions and critical feedback on the manuscript.

Appendix A. Supplementary data

The supporting information provides a detailed explanation of the methods used in this study. Moreover, it includes figures of sequence and structural alignment for EGFR and HER2 receptors, as well as the distribution of ligands according to the Lipinski, Pfizer, GSK, and golden triangle rules, along with medical synthesis parameters. Docked poses and 2D interaction views highlighting the surrounding amino acids of target proteins are also included. It also contains a table of genetic algorithm parameter values and the ranges considered for hyperparameter tuning using grid search for both active/inactive and target classifiers. Additionally, the selected features for active/inactive and target classification, as well as the evaluation results for KNN, SVM, DT, RF, NB, LDA, and QDA models for active/inactive classification and for KNN, SVM, DT, LR, RF, NB, GNB, LDA, and QDA models for target classification, are provided. These evaluations were performed using optimal parameters determined through grid search and features selected by k-best, GA-KNN, GA-GNB, GA-QDA, GA-RF, and GA-SVM methods.

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.compbiomed.2024.109279.

References

[1] F. Abedi Tameh, H.E.A. Mohamed, L. Aghababaee, M. Akbari, S. Alikhah Asl, M.H. Javadi, M. Aucamp, K.J. Cloete, J. Soleimannejad, M. Maaza, In-vitro cytotoxicity of biosynthesized nanoceria using Eucalyptus camaldulensis leaves extract against MCF-7 breast cancer cell line, Sci. Rep. 14 (1) (2024) 17465, http://dx.doi.org/10.1038/s41598-024-68272-3, URL: https://www.nature.com/articles/s41598-024-68272-3.

[2] D. Sun, W. Gao, H. Hu, S. Zhou, Why 90% of clinical drug development fails and how to improve it? Acta Pharm. Sin. B 12 (7) (2022) 3049–3062, http://dx.doi.org/10.1016/j.apsb.2022.02.002, URL: https://linkinghub.elsevier.com/retrieve/pii/S2211383522000521.

[3] F. Abedi Tameh, Z. Jahani, S. Sedghiniya, M. Amirpour Aghtaei, M. Abtahi, W. Xiang, M. Akbari, J. Soleimannejad, J. Janczak, Morphology-dependent multienzyme activity of nanoceria in antioxidant protection of MnCl2-treated PC-12 Cells, and the potential application for Parkinson's disease treatment, Inorg. Chem. Commun. 169 (2024) 113117, http://dx.doi.org/10.1016/j.inoche.2024.113117, URL: https://linkinghub.elsevier.com/retrieve/pii/S1387700324011043.

[4] M. Gackowski, K. Szewczyk-Golec, R. Pluskota, M. Koba, K. Mądra-Gackowska, A. Woźniak, Application of multivariate adaptive regression splines (MARSplines) for predicting antitumor activity of anthrapyrazole derivatives, Int. J. Mol. Sci. 23 (9) (2022) 5132, http://dx.doi.org/10.3390/ijms23095132, URL: https://www.mdpi.com/1422-0067/23/9/5132.

[5] A.V. Sadybekov, V. Katritch, Computational approaches streamlining drug discovery, Nature 616 (7958) (2023) 673–685, http://dx.doi.org/10.1038/s41586-023-05905-z, URL: https://www.nature.com/articles/s41586-023-05905-z.

[6] A. Fischer, M. Smieško, M. Sellner, M.A. Lill, Decision making in structure-based drug discovery: Visual inspection of docking results, J. Med. Chem. 64 (5) (2021) 2489–2500, http://dx.doi.org/10.1021/acs.jmedchem.0c02227, URL: https://pubs.acs.org/doi/10.1021/acs.jmedchem.0c02227.

[7] S. IsmaiL, A. UzaiRu, B. Sagagi, M.S. SuleiMan, Insilico molecular docking and pharmacokinetic studies of selected phytochemicals with estrogen and progesterone receptors as anticancer agent for breast cancer, J. Turk. Chem. Soc. Sect. A: Chem. 5 (3) (2018) 1337–1350, http://dx.doi.org/10.18596/jotcsa.449778, URL: http://dergipark.org.tr/en/doi/10.18596/jotcsa.449778.

[8] F. Shehadeh-Tout, H.H. Milioli, S. Roslan, P.J. Jansson, M. Dharmasivam, D. Graham, R. Anderson, T. Wijesinghe, M.G. Azad, D.R. Richardson, Z. Kovacevic, Innovative thiosemicarbazones that induce multi-modal mechanisms to down-regulate estrogen-, progesterone-, androgen- and prolactin-receptors in breast cancer, Pharmacol. Res. 193 (2023) 106806, http://dx.doi.org/10.1016/j.phrs.2023.106806, URL: https://linkinghub.elsevier.com/retrieve/pii/S1043661823001627.

[9] S. Gnanaselvan, S.A. Yadav, S.P. Manoharan, Structure-based virtual screening of anti-breast cancer compounds from *Artemisia absinthium* —insights through molecular docking, pharmacokinetics, and molecular dynamic simulations, J. Biomol. Struct. Dyn. 42 (6) (2024) 3267–3285, http://dx.doi.org/10.1080/07391102.2023.2212805, URL: https://www.tandfonline.com/doi/full/10.1080/07391102.2023.2212805.

[10] R. Roskoski, Properties of FDA-approved small molecule protein kinase inhibitors: A 2023 update, Pharmacol. Res. 187 (2023) 106552, http://dx.doi.org/10.1016/j.phrs.2022.106552, URL: https://linkinghub.elsevier.com/retrieve/pii/S1043661822004984.

[11] F. Skoulidis, J.V. Heymach, Co-occurring genomic alterations in non-small-cell lung cancer biology and therapy, Nat. Rev. Cancer 19 (9) (2019) 495–509, http://dx.doi.org/10.1038/s41568-019-0179-8, URL: https://www.nature.com/articles/s41568-019-0179-8.

[12] I. Balbuena-Rebolledo, I.I. Padilla-Martínez, M.C. Rosales-Hernández, M. Bello, Repurposing FDA drug compounds against breast cancer by targeting EGFR/HER2, Pharmaceuticals 14 (8) (2021) 791, http://dx.doi.org/10.3390/ph14080791, URL: https://www.mdpi.com/1424-8247/14/8/791.

[13] M. Gaibar, L. Beltrán, A. Romero-Lorca, A. Fernández-Santander, A. Novillo, Somatic mutations in *HER2* and implications for current treatment paradigms in *HER2* -positive breast cancer, J. Oncol. 2020 (2020) 1–13, http://dx.doi.org/10.1155/2020/6375956, URL: https://www.hindawi.com/journals/jo/2020/6375956/.

[14] M. Fan, M. Shan, X. Lan, X. Fang, D. Song, H. Luo, D. Wu, Anti-cancer effect and potential microRNAs targets of ginsenosides against breast cancer, Front. Pharmacol. 13 (2022) 1033017, http://dx.doi.org/10.3389/fphar.2022.1033017, URL: https://www.frontiersin.org/articles/10.3389/fphar.2022.1033017/full.

[15] G.S. Purawarga Matada, P.S. Dhiwar, N. Abbas, E. Singh, A. Ghara, A. Das, S.V. Bhargava, Molecular docking and molecular dynamic studies: screening of phytochemicals against EGFR, HER2, estrogen and NF-KB receptors for their potential use in breast cancer, J. Biomol. Struct. Dyn. 40 (13) (2022) 6183–6192, http://dx.doi.org/10.1080/07391102.2021.1877823, URL: https://www.tandfonline.com/doi/full/10.1080/07391102.2021.1877823.

[16] Z. Kalaki, M. Asadollahi-Baboli, Molecular docking-based classification and systematic QSAR analysis of indoles as Pim kinase inhibitors, SAR QSAR Environ. Res. 31 (5) (2020) 399–419, http://dx.doi.org/10.1080/1062936X.2020.1751277, URL: https://www.tandfonline.com/doi/full/10.1080/1062936X.2020.1751277.

[17] N. Singh, P. Kushwaha, A. Gupta, O. Prakash, Recent advances of novel therapeutic agents from botanicals for prevention and therapy of breast cancer: An updated review, Curr. Cancer Ther. Rev. 16 (1) (2020) 5–18, http://dx.doi.org/10.2174/1573394715666181129101502, URL: http://www.eurekaselect.com/167843/article.

[18] A. Roberti, L.E. Chaffey, D.R. Greaves, NF-κB signaling and inflammation—Drug repurposing to treat inflammatory disorders? Biology 11 (3) (2022) 372, http://dx.doi.org/10.3390/biology11030372, URL: https://www.mdpi.com/2079-7737/11/3/372.

[19] Q. Xu, J. Yu, G. Jia, Z. Li, H. Xiong, Crocin attenuates NF-κB-mediated inflammation and proliferation in breast cancer cells by down-regulating PRKCQ, Cytokine 154 (2022) 155888, http://dx.doi.org/10.1016/j.cyto.2022.155888, URL: https://linkinghub.elsevier.com/retrieve/pii/S1043466622000977.

[20] E. Pavitra, J. Kancharla, V.K. Gupta, K. Prasad, J.Y. Sung, J. Kim, M.B. Tej, R. Choi, J.-H. Lee, Y.-K. Han, G.S.R. Raju, L. Bhaskar, Y.S. Huh, The role of NF-κB in breast cancer initiation, growth, metastasis, and resistance to chemotherapy, Biomed. Pharmacother. 163 (2023) 114822, http://dx.doi.org/10.1016/j.biopha.2023.114822, URL: https://linkinghub.elsevier.com/retrieve/pii/S0753332223006121.

[21] K. Jimenes-Vargas, A. Pazos, C.R. Munteanu, Y. Perez-Castillo, E. Tejera, Prediction of compound-target interaction using several artificial intelligence algorithms and comparison with a consensus-based strategy, J. Cheminform. 16 (1) (2024) 27, http://dx.doi.org/10.1186/s13321-024-00816-1, URL: https://jcheminf.biomedcentral.com/articles/10.1186/s13321-024-00816-1.

[22] H. Shrestha, S.C.B. Jaganathan, C. Dhasarathan, K. Suriyan, Detection and classification of dermatoscopic images using segmentation and transfer learning, Multimedia Tools Appl. 82 (15) (2023) 23817–23831, http://dx.doi.org/10.1007/s11042-023-14752-z, URL: https://link.springer.com/10.1007/s11042-023-14752-z.

[23] A. Tropsha, O. Isayev, A. Varnek, G. Schneider, A. Cherkasov, Integrating QSAR modelling and deep learning in drug discovery: the emergence of deep QSAR, Nat. Rev. Drug Discov. (2023) http://dx.doi.org/10.1038/s41573-023-00832-0, URL: https://www.nature.com/articles/s41573-023-00832-0.

[24] M. Darsaraee, S. Kaveh, A. Mani-Varnosfaderani, M. Neiband, General structure-activity/selectivity relationship patterns for the inhibitors of the chemokine receptors (CCR1/CCR2/CCR4/CCR5) with application for virtual screening of PubChem database, J. Biomol. Struct. Dyn. (2023) 1–19, http://dx.doi.org/10.1080/07391102.2023.2248255, URL: https://www.tandfonline.com/doi/full/10.1080/07391102.2023.2248255.

[25] Y.-C. Lo, S.E. Rensi, W. Torng, R.B. Altman, Machine learning in chemoinformatics and drug discovery, Drug Discov. Today 23 (8) (2018) 1538–1546, http://dx.doi.org/10.1016/j.drudis.2018.05.010, URL: https://linkinghub.elsevier.com/retrieve/pii/S1359644617304695.

[26] J.A. Keith, V. Vassilev-Galindo, B. Cheng, S. Chmiela, M. Gastegger, K.-R. Müller, A. Tkatchenko, Combining machine learning and computational chemistry for predictive insights into chemical systems, Chem. Rev. 121 (16) (2021) 9816–9872, http://dx.doi.org/10.1021/acs.chemrev.1c00107, URL: https://pubs.acs.org/doi/10.1021/acs.chemrev.1c00107.

[27] V. Limongelli, Ligand binding free energy and kinetics calculation in 2020, WIREs Comput. Mol. Sci. 10 (4) (2020) e1455, http://dx.doi.org/10.1002/wcms.1455, URL: https://wires.onlinelibrary.wiley.com/doi/10.1002/wcms.1455.

[28] X. Yang, Y. Wang, R. Byrne, G. Schneider, S. Yang, Concepts of artificial intelligence for computer-assisted drug discovery, Chem. Rev. 119 (18) (2019) 10520–10594, http://dx.doi.org/10.1021/acs.chemrev.8b00728, URL: https://pubs.acs.org/doi/10.1021/acs.chemrev.8b00728.

[29] L. Pinzi, G. Rastelli, Molecular docking: Shifting paradigms in drug discovery, Int. J. Mol. Sci. 20 (18) (2019) 4331, http://dx.doi.org/10.3390/ijms20184331, URL: https://www.mdpi.com/1422-0067/20/18/4331.

[30] F. Stanzione, I. Giangreco, J.C. Cole, Use of molecular docking computational tools in drug discovery, in: Progress in Medicinal Chemistry, vol. 60, Elsevier, 2021, pp. 273–343, http://dx.doi.org/10.1016/bs.pmch.2021.01.004, URL: https://linkinghub.elsevier.com/retrieve/pii/S0079646821000047.

[31] A. Vidal-Limon, J.E. Aguilar-Toalá, A.M. Liceaga, Integration of molecular docking analysis and molecular dynamics simulations for studying food proteins and bioactive peptides, J. Agricult. Food Chem. 70 (4) (2022) 934–943, http://dx.doi.org/10.1021/acs.jafc.1c06110, URL: https://pubs.acs.org/doi/10.1021/acs.jafc.1c06110.

[32] D. Tomić, D. Davidović, A.M. Szasz, M. Rezeli, B. Pirkić, J. Petrik, V.B. Vrca, V. Janđ el, T. Lipić, K. Skala, J. Mesarić, M.M. Periša, Z. Šojat, B.M. Rogina, The screening and evaluation of potential clinically significant HIV drug combinations against the SARS-CoV-2 virus, Inform. Med. Unlocked 23 (2021) 100529, http://dx.doi.org/10.1016/j.imu.2021.100529, URL: https://linkinghub.elsevier.com/retrieve/pii/S2352914821000198.

[33] M. Awasthi, S. Singh, V.P. Pandey, U.N. Dwivedi, Molecular docking and 3D-QSAR-based virtual screening of flavonoids as potential aromatase inhibitors against estrogen-dependent breast cancer, J. Biomol. Struct. Dyn. 33 (4) (2015) 804–819, http://dx.doi.org/10.1080/07391102.2014.912152, URL: https://www.tandfonline.com/doi/full/10.1080/07391102.2014.912152.

[34] Z. Yousuf, K. Iman, N. Iftikhar, M.U. Mirza, Structure-based virtual screening and molecular docking for the identification of potential multi-targeted inhibitors against breast cancer, Breast Cancer : Targets Ther. 9 (2017) 447–459, http://dx.doi.org/10.2147/BCTT.S132074, URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5476443/.

[35] M. Anbuselvam, M. Easwaran, A. Meyyazhagan, J. Anbuselvam, H.K. Bhotla, M. Sivasubramanian, Y. Annadurai, T. Kaul, M. Pappusamy, B. Balasubramanian, Structure-based virtual screening, pharmacokinetic prediction, molecular dynamics studies for the identification of novel EGFR inhibitors in breast cancer, J. Biomol. Struct. Dyn. 39 (12) (2021) 4462–4471, http://dx.doi.org/10.1080/07391102.2020.1777899, URL: https://www.tandfonline.com/doi/full/10.1080/07391102.2020.1777899.

[36] L.K. Tsou, S.-H. Yeh, S.-H. Ueng, C.-P. Chang, J.-S. Song, M.-H. Wu, H.-F. Chang, S.-R. Chen, C. Shih, C.-T. Chen, Y.-Y. Ke, Comparative study between deep learning and QSAR classifications for TNBC inhibitors and novel GPCR agonist discovery, Sci. Rep. 10 (1) (2020) 16771, http://dx.doi.org/10.1038/s41598-020-73681-1, URL: https://www.nature.com/articles/s41598-020-73681-1.

[37] S. He, D. Zhao, Y. Ling, H. Cai, Y. Cai, J. Zhang, L. Wang, Machine learning enables accurate and rapid prediction of active molecules against breast cancer cells, Front. Pharmacol. 12 (2021) 796534, http://dx.doi.org/10.3389/fphar.2021.796534, URL: https://www.frontiersin.org/articles/10.3389/fphar.2021.796534/full.

[38] M. Aziz, S.A. Ejaz, S. Zargar, N. Akhtar, A.T. Aborode, T. A. Wani, G.E.-S. Batiha, F. Siddique, M. Alqarni, A.A. Akintola, Deep learning and structure-based virtual screening for drug discovery against NEK7: A novel target for the treatment of cancer, Molecules 27 (13) (2022) 4098, http://dx.doi.org/10.3390/molecules27134098, URL: https://www.mdpi.com/1420-3049/27/13/4098.

[39] H. Nada, A.R. Gul, A. Elkamhawy, S. Kim, M. Kim, Y. Choi, T.J. Park, K. Lee, Machine learning-based approach to developing potent EGFR inhibitors for breast cancer-design, synthesis, and in vitro evaluation, ACS Omega 8 (35) (2023) 31784–31800, http://dx.doi.org/10.1021/acsomega.3c02799, URL: https://pubs.acs.org/doi/10.1021/acsomega.3c02799.

[40] M.K. Gilson, T. Liu, M. Baitaluk, G. Nicola, L. Hwang, J. Chong, BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology, Nucleic Acids Res. 44 (D1) (2016) D1045–D1053, http://dx.doi.org/10.1093/nar/gkv1072, URL: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkv1072.

[41] N.M. O'Boyle, M. Banck, C.A. James, C. Morley, T. Vandermeersch, G.R. Hutchison, Open Babel: An open chemical toolbox, J. Cheminform. 3 (1) (2011) 33, http://dx.doi.org/10.1186/1758-2946-3-33, URL: https://jcheminf.biomedcentral.com/articles/10.1186/1758-2946-3-33.

[42] A. Frisch, Gaussian 09: User's Reference, Gaussian, Wallingford, Conn., 2009, OCLC: 711965588.

[43] A. Mauri, alvaDesc: A tool to calculate and analyze molecular descriptors and fingerprints, in: K. Roy (Ed.), Ecotoxicological QSARs, Springer US, New York, NY, 2020, pp. 801–820, http://dx.doi.org/10.1007/978-1-0716-0150-1_32, URL: http://link.springer.com/10.1007/978-1-0716-0150-1_32.

[44] O. Trott, A.J. Olson, AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading, J. Comput. Chem. 31 (2) (2010) 455–461, http://dx.doi.org/10.1002/jcc.21334, URL: https://onlinelibrary.wiley.com/doi/10.1002/jcc.21334.

[45] G. Xiong, Z. Wu, J. Yi, L. Fu, Z. Yang, C. Hsieh, M. Yin, X. Zeng, C. Wu, A. Lu, X. Chen, T. Hou, D. Cao, ADMETlab 2.0: an integrated online platform for accurate and comprehensive predictions of ADMET properties, Nucleic Acids Res. 49 (W1) (2021) W5–W14, http://dx.doi.org/10.1093/nar/gkab255, URL: https://academic.oup.com/nar/article/49/W1/W5/6249611.

[46] S. Ahmadi, A. Mani-Varnosfaderani, B. Habibi, Motor oil classification using color histograms and pattern recognition techniques, J. AOAC Int. 101 (6) (2018) 1967–1976, http://dx.doi.org/10.5740/jaoacint.17-0308, URL: https://academic.oup.com/jaoac/article/101/6/1967-1976/5654130.

[47] N. Jafarzadeh, A. Mani-Varnosfaderani, K. Gilany, S. Eynali, H. Ghaznavi, A. Shakeri-Zadeh, The molecular cues for the biological effects of ionizing radiation dose and post-irradiation time on human breast cancer SKBR3 cell line: A Raman spectroscopy study, J. Photochem. Photobiol. B 180 (2018) 1–8, http://dx.doi.org/10.1016/j.jphotobiol.2018.01.014, URL: https://linkinghub.elsevier.com/retrieve/pii/S101113441731182X.

[48] H. Ding, F. Xing, L. Zou, L. Zhao, QSAR analysis of VEGFR-2 inhibitors based on machine learning, Topomer CoMFA and molecule docking, BMC Chem. 18 (1) (2024) 59, http://dx.doi.org/10.1186/s13065-024-01165-8, URL: https://bmcchem.biomedcentral.com/articles/10.1186/s13065-024-01165-8.

[49] A.J. Siddiqui, A. Jamal, M. Zafar, S. Jahan, Identification of TBK1 inhibitors against breast cancer using a computational approach supported by machine learning, Front. Pharmacol. 15 (2024) 1342392, http://dx.doi.org/10.3389/fphar.2024.1342392, URL: https://www.frontiersin.org/articles/10.3389/fphar.2024.1342392/full.

[50] F. Bouchama, K. Kraim, M. Brahimi, Y. Saihi, K. Mezghiche, A.K. Nacereddine, A. Djerourou, M.O. Taha, Virtual screening, XGBoost based QSAR modelling, Molecular Docking and Molecular Dynamics Simulation approach to discover a new inhibitor targeting ErbB1 Protein, 2024, http://dx.doi.org/10.21203/rs.3.rs-4477079/v1, URL: https://www.researchsquare.com/article/rs-4477079/v1.

[51] D. Jiang, T. Lei, Z. Wang, C. Shen, D. Cao, T. Hou, ADMET evaluation in drug discovery. 20. Prediction of breast cancer resistance protein inhibition through machine learning, J. Cheminform. 12 (1) (2020) 16, http://dx.doi.org/10.1186/s13321-020-00421-y.

[52] A.-Q. Niu, L.-J. Xie, H. Wang, B. Zhu, S.-Q. Wang, Prediction of selective estrogen receptor beta agonist using open data and machine learning approach, DDDT 10 (2016) 2323–2331, http://dx.doi.org/10.2147/DDDT.S110603.

[53] V. Belekar, K. Lingineni, P. Garg, Classification of breast cancer resistant protein (BCRP) inhibitors and non-inhibitors using machine learning approaches, Comb. Chem. High Throughput Screen. 18 (5) (2015) 476–485.

[54] Y. Buehler, J.-L. Reymond, Molecular framework analysis of the generated database GDB-13s, J. Chem. Inf. Model. 63 (2) (2023) 484–492, http://dx.doi.org/10.1021/acs.jcim.2c01107, URL: https://pubs.acs.org/doi/10.1021/acs.jcim.2c01107.

[55] S. Kaveh, A. Mani-Varnosfaderani, M.S. Neiband, Deriving general structure–activity/selectivity relationship patterns for different subfamilies of cyclin-dependent kinase inhibitors using machine learning methods, Sci. Rep. 14 (1) (2024) 15315, http://dx.doi.org/10.1038/s41598-024-66173-z, URL: https://www.nature.com/articles/s41598-024-66173-z.

[56] J.L. Medina-Franco, A.L. Chávez-Hernández, E. López-López, F.I. Saldívar-González, Chemical multiverse: An expanded view of chemical space, Mol. Inform. 41 (11) (2022) 2200116, http://dx.doi.org/10.1002/minf.202200116, URL: https://onlinelibrary.wiley.com/doi/10.1002/minf.202200116.