www.ajlifesciences.com / ISSN: 2663-1040 *AJ Life Sci. 2026, 9 (1): 28-48*

Research Article

Abasyn Journal of Life Sciences

Open Access

DOI: 10.5281/zenodo.19657593

Title: Learning, Simulating, and Inhibiting: A Multi-Scale AI Framework for Breast Cancer Drug Discovery

Hashim M Aljohani[1,2*]

[1]Department of Clinical Laboratory Sciences, College of Applied Medical Sciences, Taibah University, Medina 42361, Saudi Arabia.

[2]Department of Pathology and Laboratory Medicine, College of Medicine, University of Cincinnati, Cincinnati, OH 45221, USA.

***Correspondence**

Department of Clinical Laboratory Sciences, College of Applied Medical Sciences, Taibah University, Medina 42361, Saudi Arabia. hsnani@taibahu.edu.sa

Abstract

Cancer is one of the most prevalent malignant diseases in the world, as breast cancer (BC) ranks as the second most frequent cause of death in women. Solubility is of great importance in drug research and development. Ensuring that molecules with the highest levels of solubility are prioritized in the initial phases of drug discovery is important in reducing the amount of resources and increasing the chances of clinical success. Herein, RF, XGBoost, LightGBM, and Artificial Neural Networks machine learning and deep learning models demonstrated over 80 percent accuracy in solubility prediction during training and testing phases. Virtual screening using a structure-based method was used against estrogen receptor alpha ligand binding domain Y537S breast cancer (BC). Best-ranked leads were selected: Hit-1 (-10 kcal/mol), Hit-2 (-9.4 kcal/mol), Hit-3 (-9.2 kcal/mol), in contrast to the control (-6 kcal/mol). ADMET (Absorption, Distribution, Metabolism, Excretion, and Toxicity) predictions were done for the top selected leads. Additionally, the binding mechanisms were assessed through molecular dynamics simulations over 150 ns. MMGBSA binding free energy of complexes was in the following order: Hit-1 (-111.34 kcal/mol), Hit-2 (-105.02 kcal/mol), Hit-3 (-112.1 kcal/mol), and control (-99.71 kcal/mol). While MMPBSA results were -107.97 kcal/mol, -103.48 kcal/mol, -107.01 kcal/mol and -97.78 kcal/mol, respectively. A dynamic cross correlation (DCCM), radial distribution function (RDF), principal component analysis (PCA), and free energy landscape (FEL) were computed, and clusters for the hit complexes were analyzed. Salt bridge interactions and secondary structure analyses indicated the identified compounds as promising binding leads. The protein-ligand complexes exhibited stable and favorable dynamics and thus can be subjected to further research to explore their breast cancer inhibition potential.

Keywords: Deep learning, ML, Breast Cancer, Molecular docking, MD simulation, MMPBSA/GBSA.

Article Info:

Received:
April 02, 2026
Received Revised:
April 13, 2026
Accepted:
April 15, 2026
Available online:
April 20, 2026

**Corresponding Author:*
hsnani@taibahu.edu.sa

1. INTRODUCTION

Uncontrolled cell division is a hallmark of a category of disorders known as cancer, which can result from a variety of variables, such as lifestyle choices, viral infections, mutations, and more, based on the GLOBOCAN estimation (Singh & Roghini, 2023). In 185 countries, about 36 different forms of cancer are diagnosed, leading to approximately 9.7 million deaths and 20 million new cases by 2050; that number is expected to increase by 77% to 35.30 million cases (Bizuayehu et al., 2024). However, with 2.3 million occurrences worldwide in 2022, breast cancer (BC) is the most often diagnosed cancer among women. BC was the fourth most common cause of cancer mortality worldwide, accounting for 666,000 fatalities (Sad, 2024). In Western nations, BC is the second leading cause of mortality for women and one of the most commonly diagnosed cancer type (Arzanova & Mayrovitz, 2022). It grows out of breast tissue (90-95%) of breast cancer cases are random ;(5-10%) are hereditary, with patients having a significant family history of the disease (ALAMUKII, 2023). About 80–90% of these inherited disease cases are caused by germ-line mutations in the BRCA1, BRCA2, and MDR1 genes (Velázquez et al., 2020). It can be roughly divided into four subtypes: triple-negative (TNBC), human epidermal growth factor receptor positive (HER2+), progesterone receptor positive (PR+), and estrogen receptor positive (ER+) (Xie et al., 2021). Any part of the breast, including the ducts, lobules, and surrounding tissues, may develop primary BC (Christgen et al., 2021). Carcinogens impact tumor microenvironments, such as stromal effects or macrophages, during metastasis, which causes BC cells to initiate and progress angiogenesis (Wu et al., 2023). Furthermore, migration to distant organs such as the liver, lung, and bone, as well as lymph nodes, is a sign of death, medication resistance, and the failure of existing treatments (Ji et al., 2023). Although radiation, surgery, and immunotherapy are the most widely used cancer treatments, their numerous side effects have driven researchers to seek less harmful alternatives (Sahu & Suryawanshi, 2021).

The nuclear hormone receptor superfamily includes the ligand-activated transcription factor known as estrogen receptor alpha (ERα) (Belluti et al., 2023). Target genes involved in normal breast development and the initiation and progression of breast cancer are either activated or repressed by estrogen binding to the ligand-binding domain (LBD) of the ER, which triggers a series of intracellular signaling cascades (Miziak et al., 2023). ERα has six structural domains, A through F, and 595 amino acids transcriptional activation function (AF)-1, which works with coregulators to control gene transcription, is present in the amino-terminal region (A/B domain) (Yu et al., 2022). The C domain has two zinc finger structures that act with DNA to regulate. ER dimers generation (Arao & Korach, 2021). The C domain is much conserved and binds to specific DNA sequences that are called estrogen response elements (EREs) (Farcas et al., 2021). The sites of phosphorylation, acetylation, and sumoylation that regulate transcriptional activity are all part of the D domain (Zhao & Malik, 2022). The protein consists of 11 alpha helices (H1, H3-12). LBD, that. 1997 The crystal structure of the first is in the E domain of ERa LBD and the natural ligand 17β-estradiol Helices 3, 4, 5, and 12 of the LBD (Lakshmanan Mangalath & Hassan Mohammed, 2021). The binding of estradiol (E2) causes the rearrangement of LBD, forming a cleft into which coactivators can bind to trigger transcription of genes (Hu et al., 2020). The F domain is made up of 45 amino acids sometimes. The so-called carboxyl-terminal domain. The partial agonist effect and tamoxifen the partial agonist property the activity of the E2-induced transcriptional activity needs to rely on the F-domain (Yu et al., 2022).

Drug solubility prediction using artificial intelligence (AI) has been of interest in the recent past as a significant alternative to check the actual outcomes of experimental studies (Tran et al., 2023). The prediction modeling and simulation of such type are developed with such methods, and many sectors (such as) are represented. The delivery, extraction, purification, and separation of drugs is able to save on computation time in a significant way assure the validity of the experimental outcomes conducted (Ponphaiboon et al., 2023). Machine learning (ML) is a collection of AI methods enabling computers to learn without being programmed (Tyagi & Chahal, 2022). To make predictions of unknown future inputs, it is necessary to train models machine learning is trying to develop meta-programs that manipulate experimentally obtained data (Wang et al., 2025). Another form of machine learning methods is ensemble approaches, which use several basic models in order to augment prediction accuracy generally (Rane et al., 2024). The pharmacokinetics, pharmacodynamics, and drug-likeness of new drug candidates were improved using computer-aided drug design (CADD) techniques (Gurung et al., 2021). In the conducted study, ML models, Random Forest (RF), XGBoost, LightGBM, and Artificial Neural Network (ANN) were employed for the sake of solubility prediction and screened against the disease by the approach of protein-ligand based virtual screening after that checked their ADMET properties and run the complexes on 150 ns

molecular dynamic simulation time frame in order to check their stability and post simulation analysis carried out as well i.e., MMPBSA/GBSA, PCA, and FEL, RDF, DCCM, secondary structure, and lastly salt bridges.

2. MATERIALS AND METHODS

Figure 1 illustrates the schematic workflow of the conducted study computationally.

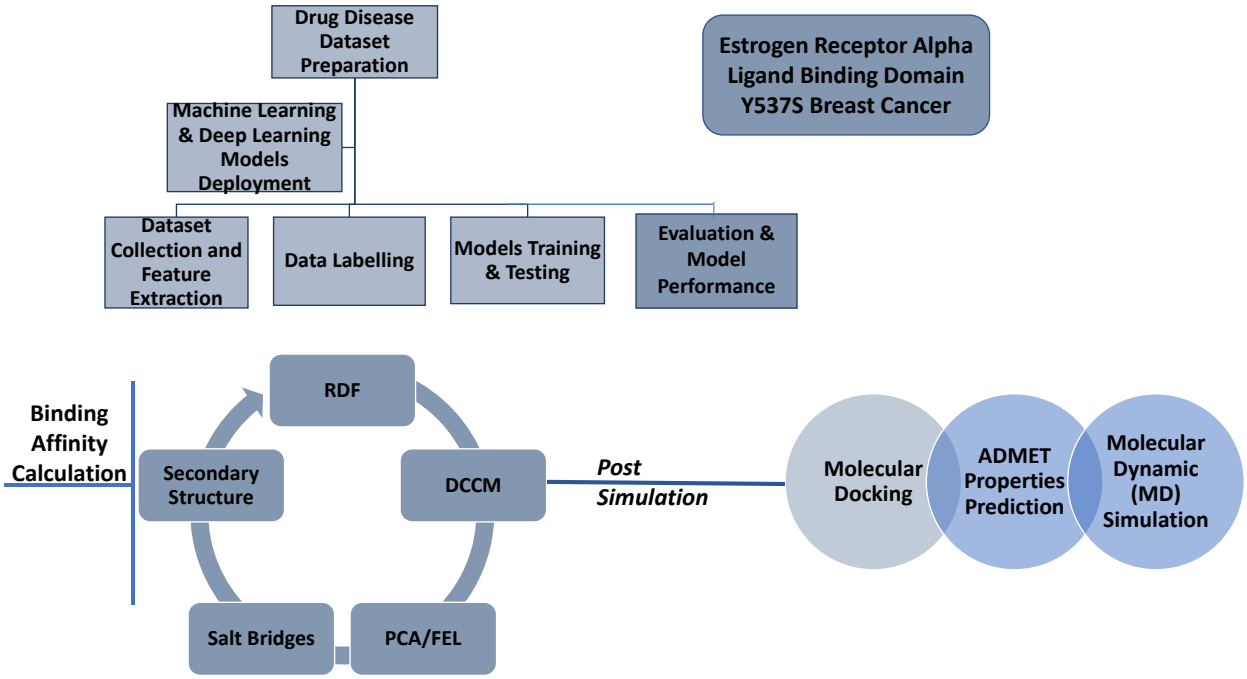


Figure 1. The illustrated workflow chart of the conducted study for the identification of novel inhibitors against Breast Cancer. The approach utilized machine learning models and a deep learning model, and their evaluation parameters during training and testing states, and confirmed through molecular docking studies, their pharmacokinetic properties, and molecular (MD) simulation studies, and post-simulation study, such as dynamic cross correlation matrix (DCCM), radial distribution function (RDF), salt bridges formation, principal component analysis (PCA) and free energy landscape (FEL) along with their calculating binding affinities respectively.

2.1. Dataset Extraction

A dataset was collected from the Plant Secondary Compounds Database (PSC-db, https://pscdb.appsbio.utalca.cl/viewSearch/index.php) of anti-cancer compounds and their respective attributes, with a total of 45 columns per compound across 1000 compounds, and was retrieved in comma-separated values (CSV) format (Islam et al., 2023).

2.1.1. Data Pre-processing

In order to allow an unbiased evaluation of the model's generalization performance, we initially split the data randomly into training and test sets with an 80:20 ratio. We then normalized the test and train data, respectively, with the aim of training and evaluating the model. Normalization of the input data is key to the effectiveness of machine learning models. The scales can be maintained with features so that some of the variables do not have a disproportionate effect on the training stage. Once the outliers were eliminated, we used Min-Max scaling to scale the features to the range between 0 and 1. The chosen scaling process is essential to the machine learning models, which depend on the magnitude of the input variables, since it normalizes the feature values so that a particular feature does not surpass other features within the period of training. Together, these preprocessing steps—removing outliers with distance and scaling features with the Min-Max scaler—improve the robustness and efficiency of our predictive models, allowing them to learn more effectively from the data and produce more accurate predictions and targeting

the feature columns of, respectively, noted most solubility attributes with the y class of *esol* and encode the target variable for further evaluation.

2.2. Machine learning models architecture

2.2.1. Random Forest (RF)

In order to increase accuracy and generalize, a machine learning technique called random forest (RF) generates a collection of aggregates and judgments based on prediction. In classification tasks, the majority of data points are typically involved, and the model reduces overfitting on a single decision tree and enjoys faster average output of multiple trees. During training, Random Forest utilizes the use of bootstrap data, where each is randomly positioned in the data and features at the time of splitting; each split improves their performance. When working with high-dimensional data, the various trees that are generated by randomness are integrated to create more accurate and dependable models. Several trees with a (n_estimators) with a maximum depth of selected tree (max_depth) with the number of trees that are essential for hyperparameters, which is used for adjustment of optimal Random Forest model performance, whereas each split with characteristics taken into the function of (max_feature) (Sandunil et al., 2024).

2.2.2. eXtreme Gradient Boosting (XGBoost)

An rXtreme Gradient Boosting (XGBoost) algorithm, a problem-solving method for supervised learning, regression, and classification models, and a popular decision-based tree integrated machine learning model. The use case of the gradient ascent framework, whether to predict a data set where ensemble learning involves developing numerous learning models of weak classifiers and integrating the projected outcome of various models into the final predicted result, is produced on a specific approach (Mienye & Sun, 2022). The convolutional gradient boosting decision tree (GBDT) algorithm is well-known for its computational efficiency, scalability, and generalization. In the context of gradient boosting, XGBoost's purpose is to minimize and normalize the loss function.

2.2.3. LightGBM (LGBM)

The decision tree serves as the foundation for the Light Gradient Boosting Machine (LGBM) framework (Hajihosseinlou et al., 2023). LGBM uses the first-order derivative information for optimizing the loss function. The leaf growth technique with depth limiting and multithread optimization in LGBM helps address excessive memory consumption compared to other boosting-ensemble machine learning methods. LGBM was adopted to reduce computational cost compared with alternative boosting ensembles.

2.2.4. Artificial Neural Network (ANN)

An artificial neural network ANN is made up of several perceptrons or neurons at each layer; it is referred to as a feed-forward neural network when the input data is sorted forward (Qamar & Zardari, 2023). The input layer, hidden layers, and output layer make up an ANN's fundamental structure. The input layer receives the input data; the hidden layers compute the input data, and the output layer offers outcomes. Each layer's responsibility in the neural networks attempts to learn precise decimal weights to be established at the end of the learning process. The ANN technique is good for tackling picture data, text data, and tabular data problems. The advantage of ANN is its ability to deal with nonlinear functions and learning weights that help map any input to the output for any data. The activation functions give the ANN nonlinear characteristics that help the net understand any complicated relationship between input and output data—a process known as a universal approximation.

2.3. Model Performance and Evaluation

A model classification performance can be evaluated by the performance assessment matrix to ensure predictions are accurate, dependable, and widely applicable. Several important parameters were often used for the following. A sensitivity of the model, which is known as true positive, analyzes how the model performed and identifies positive characteristics. It is very crucial to identify a true positive and prevent false positives in this specific field, which is renowned for its high sensitivity in illness detection. On the

other hand, the ability of the model could be accurately recognized in a negative environment, which is determined by specificity, also known as the true negative rate. In a situation where high specificity is crucial, such as spam identification and legal screening, a false positive rate may cause needless issues. Metric accuracy, or the percentage of accuracy that yielded the estimate of the overall predictions, is another frequently used metric. Whereas accuracy by its own accord sometimes can be misleading, particularly when dealing with uneven data points in a dataset. The well-known tools that are useful to separate the model performance into false positive, false negative, true positive, and true negative, allow for a more thorough model performance assessment. A model accuracy comparison bar chart was drawn for selected models with the training and testing set, along with the model fitting strategy of under- and over-fitting, as well as a feature importance plot with several features. To better comprehend model performance and ensure model effectiveness while on unseen and particular datasets, cross-validation methods like k-fold cross-validation are employed. This approach involves splitting the dataset into k-1 fragments and evaluating the model on the remaining part, whereas repeating the process of all folds, which reduces the possibility of model overfitting and generates better results with more accurate assessment throughout the model performance.

$$MCC = TP \times TN - FN \times FP / \sqrt{(TP+FN)(TP+FP)(TN+FN)(TN+FP)}$$

$$SE = TP / TP + FN$$

$$SP = TN / TN + FP$$

$$Q = TP + TN / TP + TN + FP + FN$$

2.4. Virtual Screening

An ML model for solubility class prediction, a crucial parameters were examined to better comprehend based on the accuracy of models and Matthews Correlation Coefficient (MCC). The selected phytochemicals were screened against 1000 compounds with the target protein. The major purpose is to produce and focus on a target for discovering the appealing candidate molecule that has a higher inhibitory impact on the target protein.

2.5. Molecular Docking

An Estrogen Receptor Alpha Ligand Binding Domain Y537S protein was obtained from a well-known structural database, Protein Databank (PDB), via ID 7RKE (Burley et al., 2021). Molecular docking investigations were conducted using AutoDock and AutoDock Vina, and the docking data were validated using Discovery Studio Visualizer and Visual Studio (Farooq et al., 2025). The ligand portion and water molecules from the protein structure were removed, and the hydrogen and Gasteiger charges were assigned, respectively, while the grid box was set to a 20 × 20 × 20 Å 3 box and completed using 50 GA runs per ligand. A population size of 150, a maximum of 2.5×10^2 energy evaluations, and a Lamarckian genetic technique with a spacing grid of 1 Å were the parameters of the Lamarckian Genetic Algorithm.

2.6. Pharmacokinetics Properties Assessment

Predicting the absorption, distribution, metabolism, excretion, and toxicity (ADMET) studies and the drug-likeness using the SwissADME site is an essential step in evaluating the pharmacokinetic features before proceeding to the next phase (Sucharitha et al., 2022). The top selected compounds were screened, and the following parameters were mainly focused on: physicochemical characteristics, water solubility, lipophilicity, and other pharmacokinetic properties, whereas Lipinski, Ghose, Veber, Egan, and Muegge were examined for drug-likeness, along with medicinal chemistry evaluation (Pantaleão et al., 2022).

2.7. Molecular Dynamic (MD) Simulation

An AMBER 22 molecular dynamic (MD) simulation of complex structural stability and protein-ligand interactions with dynamic behaviour throughout a 150 ns simulation period (Shukla & Tripathi, 2020). Protein ligand docking studies were applied to determine the ligand binding pocket under a static

environment (Kapoor et al., 2022). There is also the possibility of predicting the ligand binding site in physiological conditions with the help of the MD simulation (Grewal et al., 2025). The selected complexes were first processed using the Python script. The chemicals were designed with the help of the force field 5 GAFF, and a receptor protein was designed with the FF14Sb force field (He et al., 2020). The three key steps in the MD simulation are the creation of prmtop, the preparation step, and the production phase (Matamoros-Recio et al., 2023). First, the TIP3P water box was added when counter ions were added. A selected complex was exposed to a 150 ns time frame generation and progressively heated to 310 K, and given time for equilibration (Mosa et al., 2024). The complex 57 structure-based screening was evaluated and investigated using the CPPTRAJ. A Python script was used for plot generation in order to create the simulation's graphs.

2.8. Calculating Binding Affinities MMPBSA/GBSA

Examining the interaction between ligands and proteins is necessary, and it requires an understanding of free energy binding calculation research (King et al., 2021). The complex's free energy (ΔG binding) was determined using the AMBER22 package's MM/PBSA technique. In contrast, the primary enzyme's binding free energy was calculated using MMPBSA (Gogoi et al., 2021). Binding energies were analyzed using the following formulation:

$$EMM=\Delta Eint+\Delta Eele+\Delta Evdw$$

$$\Delta Gasol = \Delta Gp+\Delta Gnp$$

$$\Delta Gtotal = \Delta EMM+\Delta Gsol$$

$$\Delta Gbind = \Delta EMM+\Delta Gsol-T$$

The symbols have different meanings. For instance, ΔEele represents electrostatic energy, ΔEint represents internal interaction energy, and ΔEMM represents the total change in molecular mechanics energy. Additionally, ΔEvdw stands for the van der Waals energy change, ΔGp for the total polar solvation energy change, and ΔGnp for the non-polar solvation energy change.

2.9. Principal Component Analysis (PCA) and Free Energy Landscape (FEL)

A principal component analysis (PCA) is an approach by which the different motional fluctuations in the protein structure over the manufacturing period, which are essential dynamics (EDs) (Moradi et al., 2024). The PCA and FEL analysis through a method of covariance with the enzyme Cα atoms that were modified by SMT on the aforementioned formulation:

$$C=< (qi - <qi>) (qj - <qj>) T>$$

Conversely, it displays the mean locations of the ith and jth Cαatoms in the protein of interest with regard to structural formations obtained from MD simulation, respectively. When coupled, the diagonalization-produced eigenvalue and eigenvector demonstrate the coordination of the structural domain's movements and the variation in their intensities along an eigenvector. The Python package was distributed by a PCA.

2.10. Radial Distribution Function (RDF)

The radial distribution function (RDF), denoted by $g\ (r)$, offers the possibility of acquiring a pair of atoms within the designated range of radial detachment during MD simulation (Lamichhane & Ghimire, 2021). The equation defines the normalized RDF for atom pairs j in a spherical shell with radius r and thickness dr.

$$gj,k(r) = \frac{Nj,k(r)/4\pi x^2}{\sum r(Nj,k(r)/4\pi x^2)}$$

$Nj,(r)$ represent the number of occurrences of atom pairs that appear between r and $r+drj,k(r)$.

2.11. Dynamic Cross Correlation Matrix (DCCM)

Dynamic Cross-Correlation matrix (DCCM) investigations were used to estimate the correlation matrix for specific complexes in order to differentiate Cα (Salehi & Meuwly, 2022). A DCCM usually consists of two types of graphs: positive and negative correlations. While the movement of protein and ligand follows the same direction, the protein creates a stable relationship that comes in positive correlation. As a result, there is a negative correlation and instability when the ligand shifts away from the protein binding pocket. The color intensity in the DCCM map indicates the strength of positive and negative correlations.

2.12. SS Evaluation

Secondary structure analysis was conducted in order to establish any variations in the patterns of secondary structures of the selected proteins (Sadat & Joye, 2020). Submission of the targeted complexes to the secondary structural analysis was performed with the help of a Python .in script.

2.13. Salt Bridges Formation

An assessment of salt bridges was carried out through a Python package with the salt.in script for numerous reasons, salt bridge analysis is essential in computational drug design and is carried out with salt. Salt bridges between oppositely charged amino acid residues stabilized protein structures and influenced molecular interactions in a protein (Mutharasappan et al., 2020). Researchers can uncover critical connections necessary for drug binding by using computational analysis of salt bridges to understand the stability and dynamics of protein-ligand interactions (Klebe, 2025).

3. COMPUTATIONAL RESULTS

3.1. Evaluation of Machine Learning Models

In the carried out study, we used the phytochemical dataset that was downloaded into the CSV format of the PSC-db database and loaded into the Google Colab repository to run the pipeline of multiclass solubility prediction, where the RDKit module and Scikit-learn, along with other libraries, including Pandas, NumPy, and Matplotlib, were used. The models applied were the following: Random Forest (RF), XGBoost, LightGBM, and Artificial Neural Network (ANN) to determine the number of compounds that are soluble and insoluble.

The data was divided into training and testing data set with the concept of 80: 20 proportion. The model performance was evaluated and making ensure the prediction accuracy of the models was examined, and the parameters for each algorithm. Metrics like Q+ stand for positive predictive value (precision), Q− for negative predictive value, SP for specificity (false positive rate), and SE for sensitivity (true positive rate) were used to evaluate the model's performance. ACC: precision; Matthews' correlation coefficient, or MCC, is shown and visualized along with model overfitting and under fitting strategies with the cross-validation of performance and confusion matrix that is shown in **Figure 2 (A, B, C, D, E & F)** and **Table 1,** respectively. The model's performance was sufficient to rely on the predictions, as all models exceeded 80% accuracy during training and testing. During training, the recorded accuracy was 94% for RF, 94% for XGBoost, 98% for LightGBM, and 86% for Neural Network. On the testing set, the accuracy was 91%, 93%, 94%, and 87%, respectively. The overfitting assessment indicated a minimal overfitting gap for the top models with Cross-validation results further confirming which showed robustness and stability of the ensemble methods. Whereas Feature-importance analysis highlighted lipophilicity-related descriptors as key contributors. The LightGBM confusion matrix demonstrated strong class-wise insight, particularly for soluble and very soluble compounds categories.

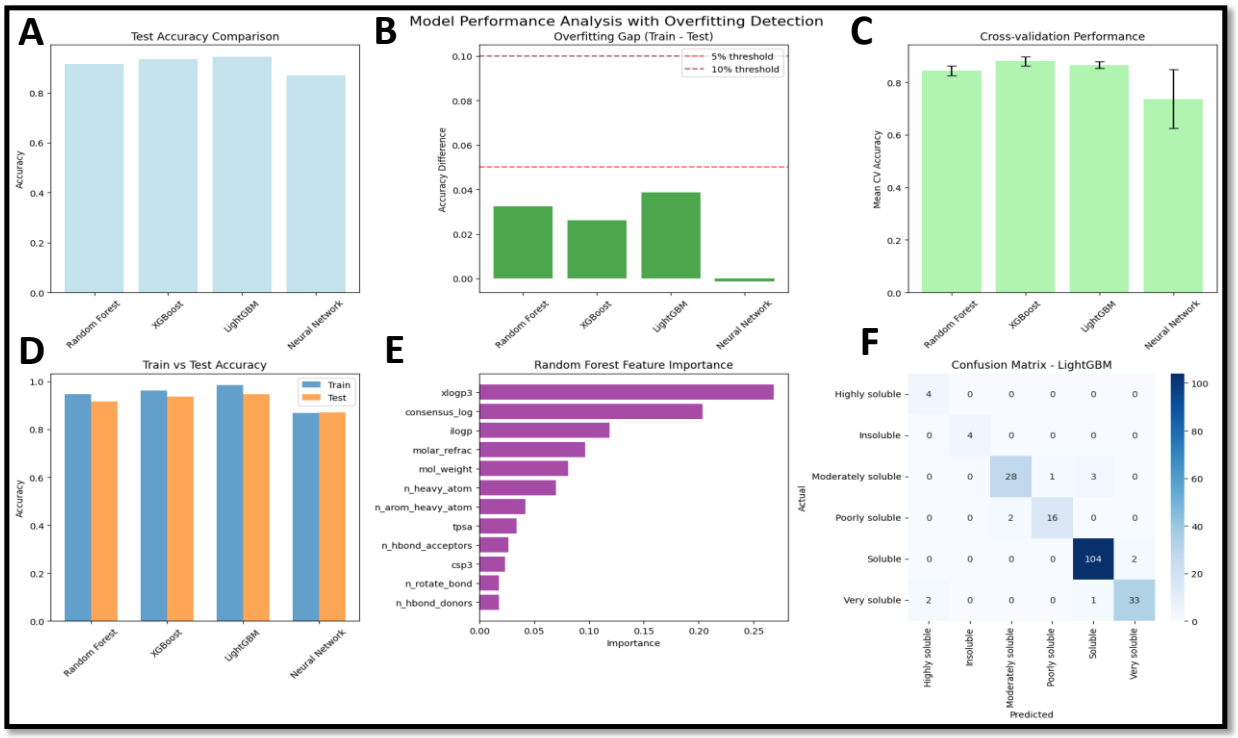


Figure 2. A comprehensive model evaluation for solubility prediction through ML models **(A)** test accuracy comparison across Random Forest, XGBoost, LightGBM, and Neural Network **(B)** an assessment of model overfitting and threshold gap for training and testing set gap with the reference **(C)** cross-validation of performance with mean accuracy and side bar **(D)** side-by-side comparison of train and test set accuracy for selected models **(E)** Feature importance ranking from the Random Forest classifier highlighting the most influential molecular descriptors and lastly **(F)** Confusion matrix of the LightGBM model illustrating classification performance across solubility categories.

Table 1. Hyperparameters of models, such as **Q+** stands for positive predictive value (precision); **Q–** for negative predictive value; **SE** for sensitivity (true positive rate); and **SP** for specificity (false positive rate). **ACC**: precision; The Matthews' correlation coefficient, or **MCC,** for Random Forest (RF), XGBoost, LightGBM, and Neural Network during training and testing mode.

Dataset	Model	SE	SP	Q+	Q-	ACC	F1 Score	MCC
Testing Set	Random Forest	0.82	0.97	0.90	0.98	0.91	0.84	0.86
	XGBoost	0.93	0.98	0.88	0.98	0.93	0.90	0.90
	LightGBM	0.94	0.98	0.90	0.98	0.94	0.91	0.91
	Neural Network	0.63	0.97	0.90	0.98	0.87	0.62	0.79
Training Set	Random Forest	0.90	0.98	0.97	0.98	0.94	0.93	0.91
	XGBoost	0.93	0.98	0.97	0.98	0.94	0.93	0.91
	LightGBM	0.96	0.99	0.98	0.99	0.98	0.97	0.97
	Neural Network	0.68	0.96	0.70	0.97	0.86	0.69	0.79

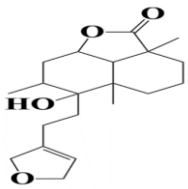
3.2. Virtual Screening

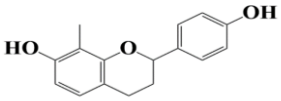
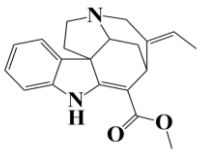
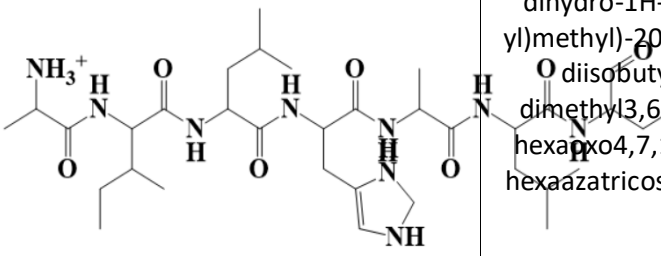
Using 1000 phytochemical anti-cancer compounds from the PSC-db database, a virtual screening was performed to determine and evaluate the efficacy of the selected dataset and practical approach. As an output of the screening procedure, compounds were subjected to a molecular docking approach to assess their binding ideal position within the active pocket and interaction potential of the protein having ID 7RKE against breast cancer.

3.3. Molecular Docking Analysis

The study objective is to target breast cancer and identify a novel inhibitor that slows down the activity of the protein. The obtained PDB structure with ID 7RKE, an Estrogen Receptor Alpha Ligand Binding Domain enzyme with a total of 547 amino acid residues, consists of an A and B chain, where chain A is removed. The compounds were then docked against chain B, utilizing active site residues. The Chimera v.1.18 was used to visualize the structure. **Figure 3 (A, B, C & D)** illustrates the structure before the docking step, with a close-up of the active pocket. For molecular docking, the protein structure was produced using UCSF Chimera, which eliminated any bound ligands, heteroatoms, and water molecules to make the structure clean. 1000 phytochemical compounds were screened against the target protein using PyRx 0.8 for docking studies using the AutoDock vina program. Compounds were then identified based on binding score. **Table 2** demonstrates, and is coupled with their chemical name, structure, binding affinity, and their IDs. In order to verify the docking results and assess their potential as breast cancer enzyme inhibitors, three of the ten compounds were chosen for additional computational investigation, particularly molecular dynamic modeling. The high-throughput screening showed the most prominent complexes as **Hit-1** 6(2(2,5dihydrofuran-3-yl)ethyl)-6-hydroxy-2a,5a,7-trimethyldecahydro-2H-naphtho[1,8-bc]furan-2-one, **Hit-2** 2-(4hydroxyphenyl)-8-methylchroman-7-ol, Hit-3 (E)methyl12ethylidene-1,2,3a,4,5,7hexahydro-3,5-ethanopyrrolo[2,3-d]carbazole6carboxylate and **Control** as 5-(sec-butyl)-11-((2,3-dihydro-1H-imidazol-4-yl)methyl)-20-formyl-8,17-diisobutyl-14,22-dimethyl3,6,9,12,15,18-hexaoxo4,7,10,13,16,19-hexaazatricosan2aminium with the binding affinity of -10 kcal/mol, -9.4 kcal/mol, -9.2 kcal/mol and -6 kcal/mol respectively. **Figure 3** illustrates the binding interaction with the residues shown as a 2D pose of complexes, while **Hit-1** revealed the highest binding affinity of -10 kcal/mol with the interaction among the residues van der Waals Met528, Met343, Thr347, Leu346, Glu353, Gly521, Ile424, Gly390, and hydrogen bonds His524, Leu525, Leu387, Arg394, while Pi-Sigma and Pi-Alkyl Leu384, Met388, Ala350, Leu391, Leu394, while **Hit-2** having the second highest binding affinity score -9.4 kcal/mol with the interaction residues of van der Walls of Met388, Phe404, Leu428, Met343, Met421, Gly521, Met528, Glu353, Arg394, Leu349, Leu346 hydrogen bonds Leu387, while Alkyl, Pi-Alkyl, Pi-Pi-T-shaped Leu391, Ile424, Leu525, Leu384, and His524 **Hit-3** the third highest binding affinity -9.2 kcal/mol with the residue interaction of van der Walls Met528, Met343, Thr347, Leu387, Ala350, Leu349, Ile424, Met421, His524 lastly the **Control** complex with the binding affinity of -6 kcal/mol with an interaction of van der Walls Asn519, Thr460, Glu385, Ile452, Asn455 with the hydrogen bond Ser456, Ser512 via Pi-Sigma, Pi-Alkyl, and Alkyl Leu508, Leu511, Arg515, and Tyr459.

Table 2. The top hit selected complexes from high-throughput screening with their binding affinities in kilocalories per mole (kcal/mol), having their structure, and compound name, respectively.

S.No	Compound ID	Compound Structure	Compound Name	Binding Affinity
1	**Hit-1**		6(2(2,5dihydrofuran-3-yl)ethyl)-6-hydroxy-2a,5a,7-trimethyldecahydro-2H-naphtho[1,8-bc]furan-2-one	-10 kcal/mol

2	**Hit-2**		2-(4hydroxyphenyl)-8-methylchroman-7-ol	-9.4 kcal/mol
3	**Hit-3**		(E)methyl12ethylidene-1,2,3a,4,5,7hexahydro-3,5-ethanopyrrolo[2,3-d]carbazole6carboxylate	-9.2 kcal/mol
4	**Control**		5-(sec-butyl)-11-((2,3-dihydro-1H-imidazol-4-yl)methyl)-20-formyl-8,17-diisobutyl-14,22-dimethyl3,6,9,12,15,18-hexaoxo4,7,10,13,16,19-hexaazatricosan2aminium	-6 kcal/mol

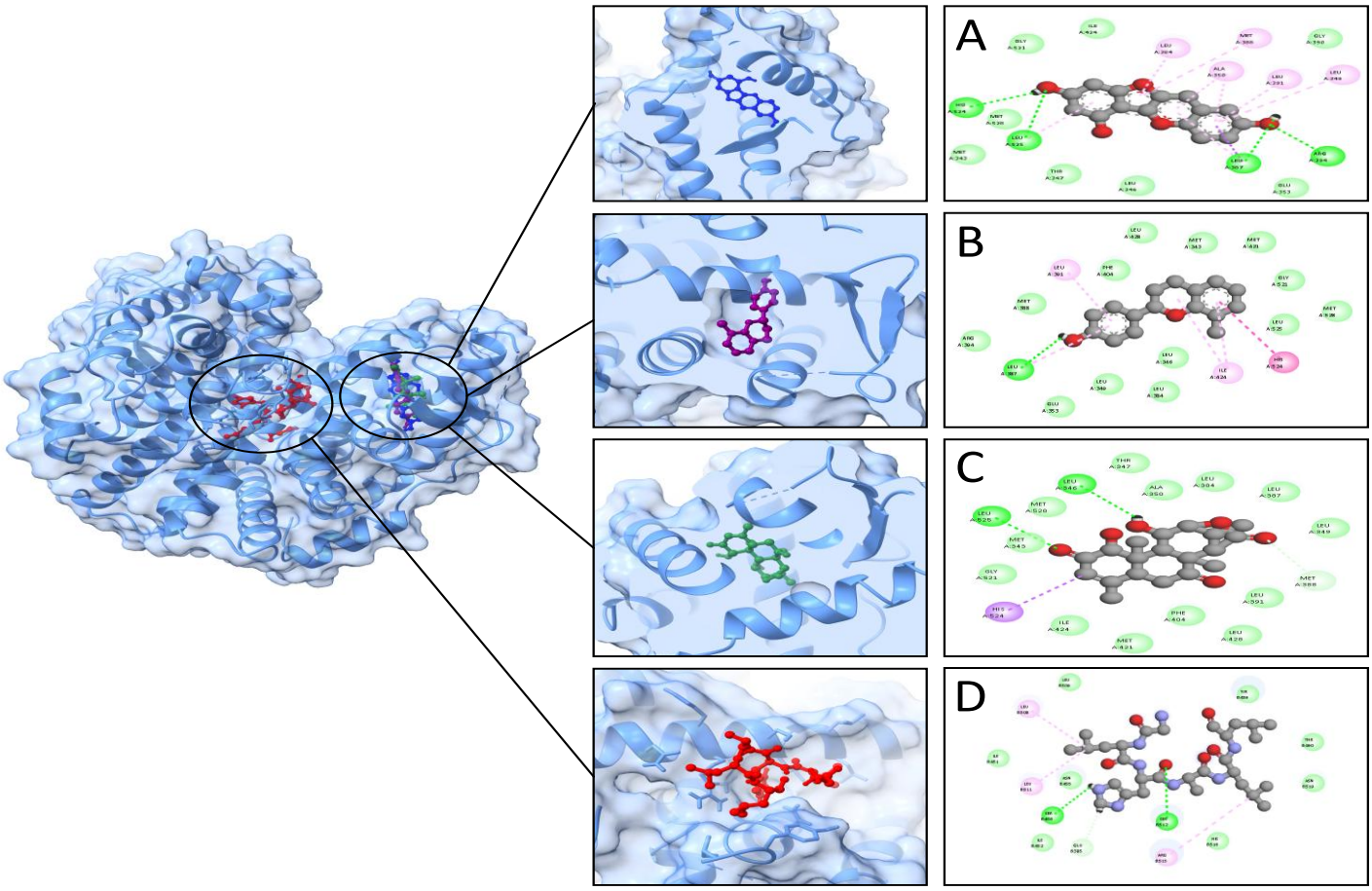


Figure 3. A molecular docking output of the top hit complexes as Hit-1 (A), Hit-2 (B), Hit-3(C), and Control (D), as shown in 3D and 2D poses with the target protein against BC within the active pocket of the hit complexes, along with their interactions, respectively.

3.4. Absorption, Distribution, Metabolism, Excretion (ADMET) Assessment

The ADMET characteristics of every ligand that binds to the receptor protein were evaluated using a SwissADME web server (Sardar, 2023), which is summarized in **S-1 Table**. Examined were the compounds' physicochemical properties, such as their molecular weight, chemical formula, percentage of Csp3, number of aromatic heavy atoms, TPSA, number of H-bonds, acceptors, and donors. The lipophilicity, pharmacokinetics, and additionally, the compounds' water solubility were assessed. The pharmacokinetic profile analysis was then used to examine the compounds' GI absorption, drug-likeness, BBB saturation, P-gp substrate, and medicinal chemistry properties. The substances were determined to be drug-like chemicals. due to their acceptable physicochemical features and obey the rule strategy. Furthermore, it was demonstrated that the drugs' pharmacokinetic range, oral bioavailability, and GI absorption were adequate. Because of their high synthetic accessibility score, the compounds are easier to synthesize. Additionally, the compounds' effects were specific because the PAINS evaluation for the chosen complexes showed no alarms for Hit-1, Hit-2, Hit-3, and the Control compound did not follow the drug likeness rule.

3.5. Molecular Dynamic (MD) Simulation Analysis

During a molecular dynamic (MD) simulation, the docked ligand sites' consistency and stability were assessed. Root Mean Square Deviation **(RMSD)**, Radius of Gyration **(Rg)**, Beat factor **(β-factor)**, Solvent Accessible Surface Area **(SASA)**, and Ligand RMSD data were generated by the trajectories. For 150 ns, the top-selected docked complexes were simulated (Li et al., 2020).

The stability of the protein–ligand interaction during the simulation is shown in the RMSD plot. A continuously rising RMSD indicates structural deviance, while a plateau indicates balance (Sasidharan et al., 2023). The lead complexes Hit1, 2, 3, along with the control molecule, the .dat files were subjected to Python v3.13 for the sake of plotting RMSD, RMSF, Rg, B-factor, SASA, and ligand-RMSD beside through

distribution respectively. The RMSD Hit-1 (Green) with the minimum of 0.0 Å values, with an average of 1.97 Å, and with a maximum of 3.13 Å through the simulation period, while Hit-2 (Brown) contrasts with the minimum of 0.0 Å, with the average of 1.67 Å, and a maximum of 2.51 Å examined while Hit-3 (Gray) is still with minimum of 0.0 Å with an average of 1.68 Å and maximum of 2.54 Å lastly the Control (Blue) molecule is with minimum value of 0.0 Å with an average of 1.78 Å and maximum of 2.38 Å respectively and the distribution for each hit can also be examined with the parallel graph which is shown in **Figure 4 (A, B)**. The flexibility of amino acid residues in protein–ligand interactions is revealed by RMSF analysis. Peaks in the RMSF are thought to signify more flexible regions, whereas troughs are assumed to suggest rigid regions (Zhuo et al., 2023). The RMSF leads Hit-1 (Purple) demonstrated within the time frame of 150 ns where the minimum values noted as 0.48 Å with an average of 1.09 Å and maximum of 8.03 while Hit-2 (Pink) also exhibited a minimum value of 0.43 Å closely similar to Hit1 (Purple) with an average of 0.95 Å and maximum of 5.23 Å eventually the Hit-3 (Yellow) exhibited the values with minimum of 0.44 Å an average of 0.96 Å and maximum was 4.72 Å whereas the Control (Red) noted with the minimum of 0.47 Å average value of 0.96 and the maximum 4.80 Å which can be seen in **Figure 4 (C, D)** long with distribution plot.

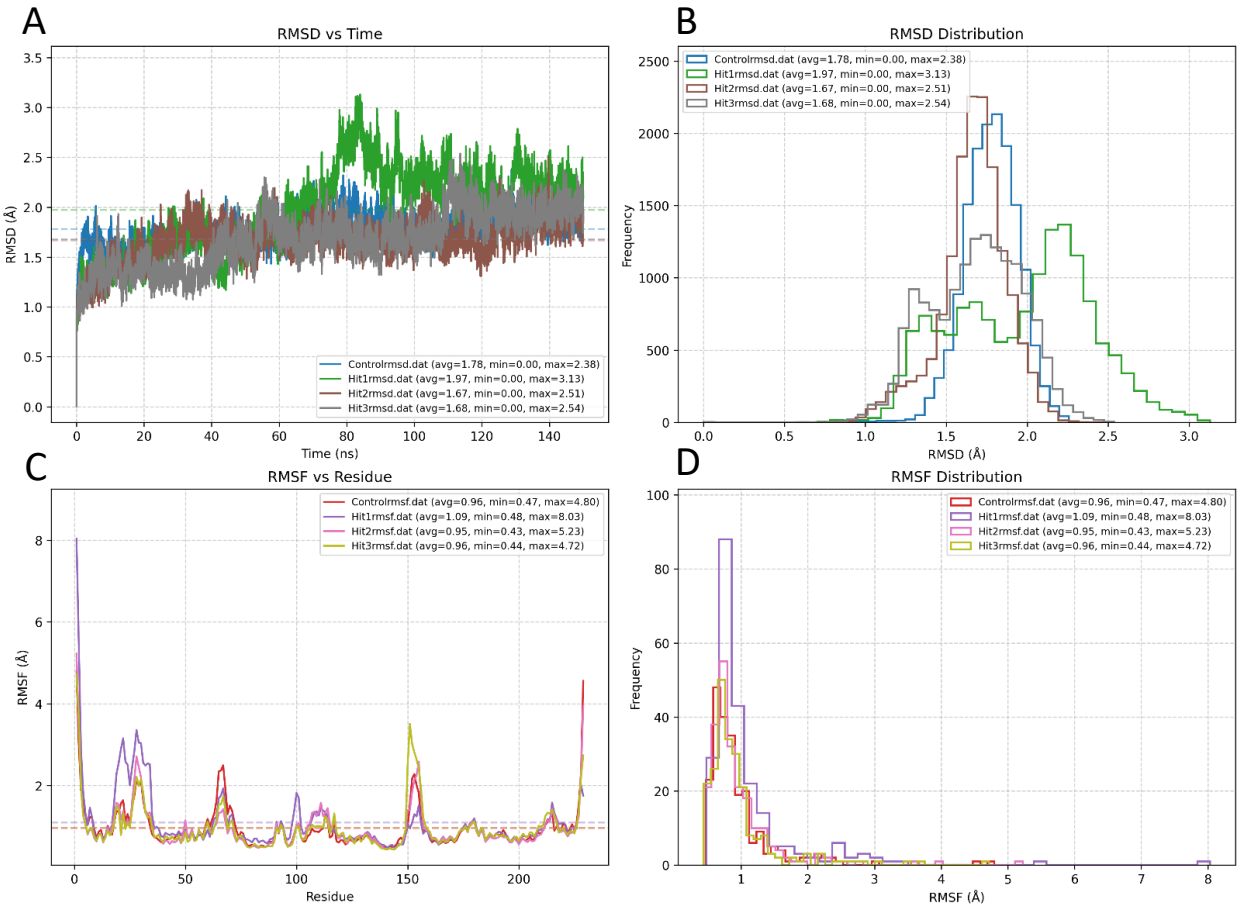


Figure 4. The RMSD along with distribution **(A, B)** and the RMSF with distribution **(C, D)** of top lead complexes Hit-1, Hit-2, Hit-3, and Control molecule within the 150 ns simulation time frame, including their minimum, average, and maximum score.

The radius of gyration (Rg) is used to check the protein-ligand behavior during the simulation period and identify rigid regions. Rg indicates expansion or unwinding as it rises, and compactness and rigidity when it falls (Dohnalová et al., 2024). According to the simulation time frame, each complex showed stability and structural compactness; Hit-1 (Red) showed the minimum value of 18.50 Å with an average of 18.93 Å and a maximum of 19.41 Å, while Hit-2 (Purple) showed the most similar behavior to the previous, with a minimum of 18.51 Å with an average of 18.90 Å and maximum of 19.48 Å whereas the Hit-3 (Pink) showed slightly similar output as previously explained molecules with minimum of 18.49 Å an average of 18.88 Å

and with maximum of 1933 Å respectively lastly the Control (Black) exhibited with minimum of 18.57 Å with the average of 18.98 Å and maximum of 18.98 Å which can be seen along with the frequency distribution in **Figure 5 (A, B)**. The protein-ligand complex mobility and flexibility of atoms can be examined by analyzing the β-factor during the simulation period (Aldakheel & Alduraywish, 2025). The β-factor for Hit-1 (Blue) indicates the minimum values of 6.01 Å average of 47.41 Å, and a maximum of 1698.26 Å, while Hit-2 (Green) revealed the minimum of 4.94 Å, with an average of 31.84 Å, with a maximum of 718.68 Å, and Hit-3 (Brown) examined with the minimum of 5.04 Å, with the average of 32.52 Å and maximum of 585.87 Å respectively. At the same time, the Control molecule exhibited with minimum of 5.85 Å with an average of 32.10 Å and a maximum of 606.05 which can be seen along with frequency distribution in **Figure 5 (C, D**).

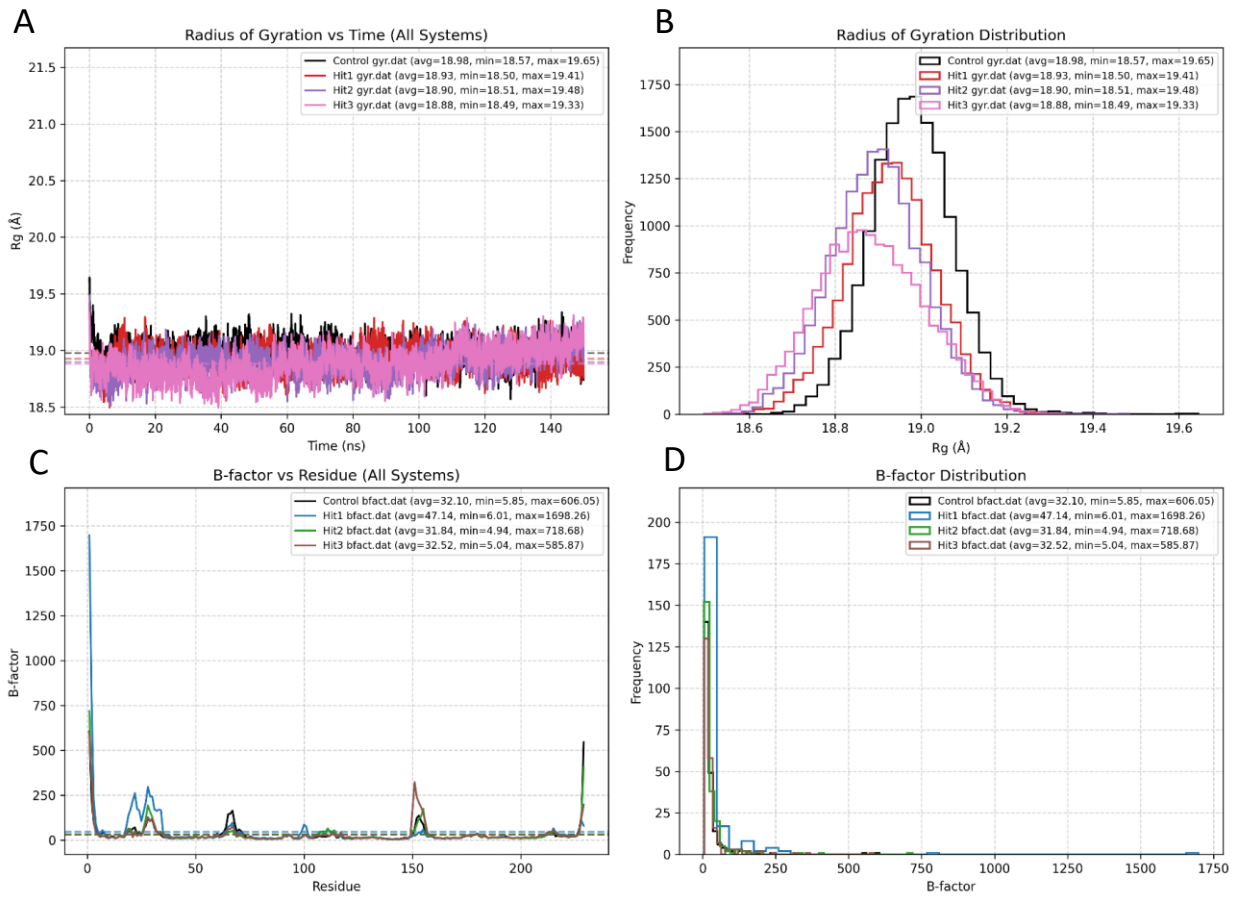


Figure 5. A depiction of Rg **(A, B)** and β-factor **(C, D)** during a simulation period of 150 ns, along with their frequency distribution estimation of selected complexes with their minimum, average, and maximum values.

3.6. Solvent Accessible Surface Area (SASA) and Ligand RMSD Assessment

The confirmation of solvent molecules interacting with the target protein's surface, the top hit complexes go through a solvent-accessible surface area (SASA) analysis. The selected complexes were examined as Hit-1 (Red) with a minimum of 11924.36 Å^2 an average value of 1308.55 Å^2 with a maximum of 14408.12 Å^2 while Hit-2 (Purple) peaked with the minimum of 11685.16 Å^2 an average of 12826.75 Å^2 and maximum of 13800.12 Å^2 whereas Hit-3 (Pink) having a minimum of 11806.23 Å^2 with an average of 13098.89 Å^2 with a maximum of 14488.02 Å^2 eventually the Control (Black) complex starting from the minimum of 12506.75 Å^2 with an average of 13497.97 Å^2 with a maximum of 14643.73 Å^2 which can be seen and illustrated along frequency distribution in **S-Figure 1 (C, D)**. At the same time, Ligand-RMSD for the selected complexes was evaluated during the simulation period in order to note the rotational, along with translational variations. The Hit-1 (Blue), Hit-2 (Green), Hit-3 (Brown), and Control (Black), the entire complexes having a minimum value of 0.0 Å^2 with an average of 1.97Å^2, 1.67 Å^2, 1.68 Å^2, and 1.78 Å^2, with a maximum of 3.12 Å^2, 2.50 Å^2,

2.54 Å^2, and 2.38 Å^2 while the depicted illustrated along with frequency distribution in **S-Figure 1 (A, B)** respectively.

3.7. MMPBSA/GBSA Analysis

By utilizing an Amber22 package, the analysis of MMBPSA/GBSA takes place and estimates the binding free energy of the top docked selected complexes, as noted Hit-1, Hit-2, Hit-3, and Control molecule, and is shown in **Table 3**. The net energy was examined as -111.34 kcal/mol, -105.02 kcal/mol, -112.1 kcal/mol, and Control -99.71 kcal/mol, while the MMPBSA were noted as -107.97 kcal/mol, 103.48 kcal/mol, 107.01 kcal/mol, and Control -97.78 kcal/mol, respectively. Due to their high net binding energy scores, the molecules in these complexes are predicted to form robust and durable intermolecular interactions.

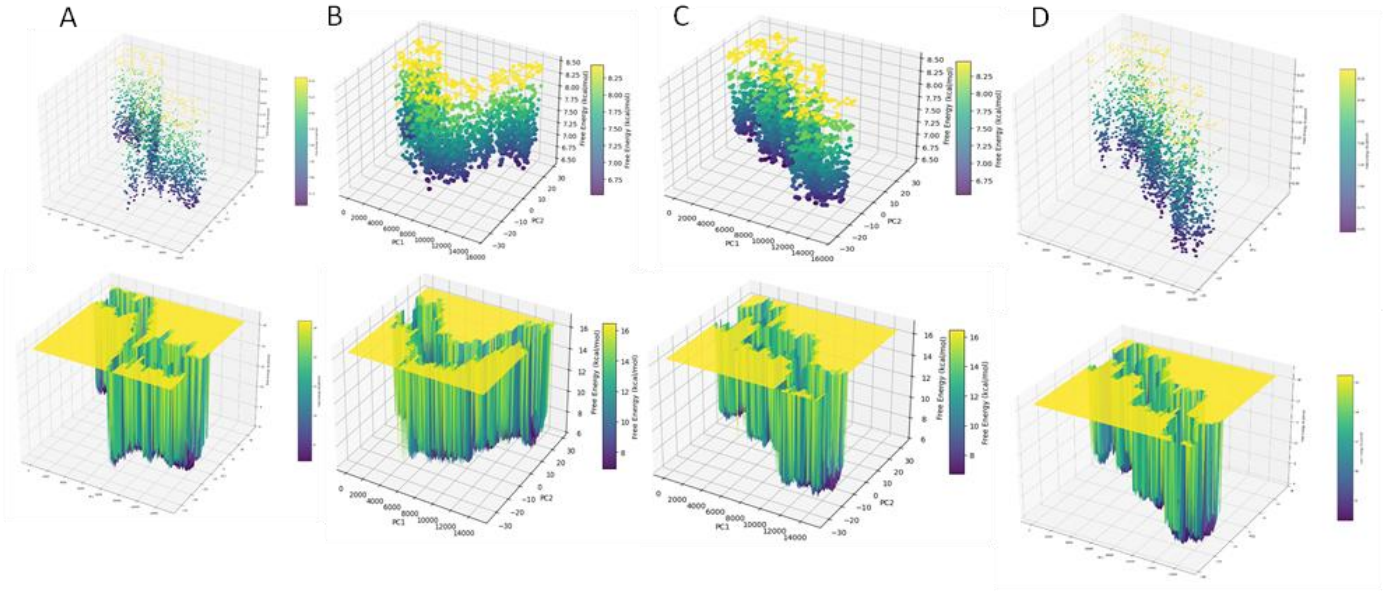
Table 3.MMGBSA/PBSA investigation of particular complexes, their control molecule, and their net energy parameters, respectively.

Energy Parameters	Hit-1	Hit-2	Hit-3	Control
MMGBSA				
Van der Waals energy	-101.30	-98.03	-100.97	-95.30
Energy electrostatic	-26.34	-24.00	-27.13	-19.48
Total gas phase energy	-127.64	-122.03	-128.1	-114.78
Total salvation energy	16.30	17.01	16.00	-15.07
Net energy	-111.34	-105.02	-112.1	-99.71
MMPBSA				
Energy van der Waals	-101.30	-98.03	-100.97	-95.30
Energy electrostatic	-26.34	-24.00	-27.13	-19.48
Total gas phase energy	-127.64	-122.03	-128.1	-114.78
Total energy salvation	19.67	18.55	21.09	17.00
Net energy	-107.97	-103.48	-107.01	-97.78

3.8. Principal Component Analysis (PCA) and Free Energy Landscape (FEL)Assessment

The 2D projection of the MD trajectories on eigenvectors 1 and 2 (PC1 and PC2) was emphasized. Over the course of the investigation, these PCs showed the most structural variance among the complexes. **Figure 6** Hit-1 (A), Hit-2 (B), Hit-3 (C), and Control (D) shows the structural variation for each of the complexes selected at the top. In this case, higher graph dispersion indicates significant conformational diversity. Each point on the graph denotes a distinct conformation, and every node represents a unique confirmation and complex stability with varying degrees of flexibility for all the systems. Whereas (B) and (C) indicate broader

distribution that suggest greater conformational stability, while (A) and (D) exhibit more localized points of



conformations, which reflect restricted and stable conformations.

Figure 6. Free energy landscape (FEL) volcano plots and principal component analysis (PCA) density plots for the chosen complexes as Hit-1 (A), Hit-2 (B), Hit-3 (C), and Control (D).

3.9. Dynamic Cross Correlation Matrix (DCCM) Analysis

The dynamic cross correlation matrix (DCCM) shows the interaction of protein ligand formation was computed for specific complexes and shows a notably large-scale positive correlation (red regions), whereas the molecules found a stable positive correlation in the binding site and the (blue) anti-correlated negative ranges and flexibility regions link to functional conformational changes shown in **Figure 7 (A, B, C & D)** Control, Hit-1, Hit-2, and Hit-3 respectively. Where the Control showed strong diagonal interaction around atom index 200-350 and 700-900, while Hit-1 showed the improved clusters around 150-300, 600-850, and 1000-1200, Hit-2 indicates the strongest interaction among the complexes in the range of 200-400, 650-900, and 950-1200 lastly the Hit-3 predicted with a broad regions in 250-450, 700-950, and 1000-1200 whereas the Hit-2 and Hit-3 showed the stronger correlated and anti-correlated interactions which indicate the stability of complexes and structural communication.

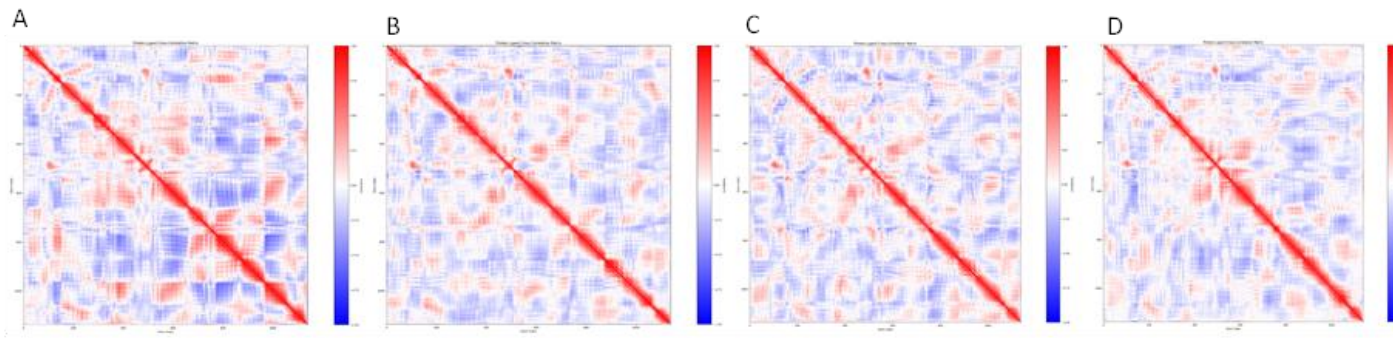


Figure 7. A dynamic cross-correlation matrix (DCCM) of the selected complexes: Control (A), Hit-1 (B), Hit-2 (C), and Hit-3 (D), with negative (blue) and positive (red) correlations examined across residues.

3.10. Radial Distribution Function (RDF) Analysis

The radial distribution function (RDF) analysis looks at how density changes at the protein complex interface as interacting residues get beyond apart. Displays a peak in all complexes, including Control. A

specific interaction between the protein and ligand at the junction is strongly shown by a single peak. A protein's stability with a chemical over time is shown by an RDF. All the complexes rise around the peak of ~ 3 to 4 Å with significant molecular interactions. The **Figure 8** Hit-1 (A) shows around ~ 3 Å with reaching a maximum of ~ 0.012, while Hit-2 (B) indicates and rise after ~ 3.5 Å and reaches around ~ 0.011, suggesting a stable but lighter, weaker interaction, whereas Hit-3 (C) reaches the highest coordination, peaking at ~ 0.014 close to 9-10 Å having string and more ordered interactions lastly Control (D) with the smooth increase in in range of ~ 0.013 with a moderate level of interaction.

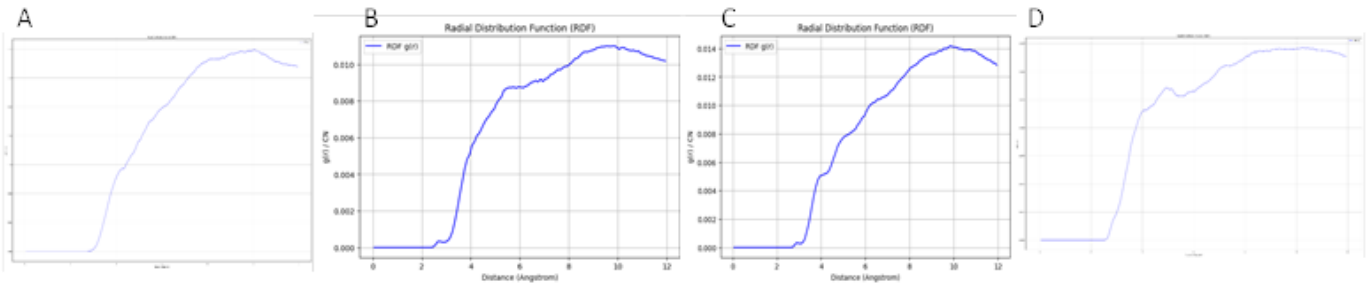


Figure 8. An illustrated graph of the Radial Distribution Function (RDF) for top selected protein-ligand complexes against BC as Hit-1 (A), Hit-2 (B), Hit-3 (C), and Control (D) molecule, respectively.

3.11. Secondary Structure Formation

A well-known beta sheet and Alpha helices, two secondary structural components that provide complex stability and include the active sites necessary for enzymatic action, constitute the overall structure of the protein-enzyme complex during analysis (Niazi, 2025). Tight turns and loops in the enzyme component aid in substrate binding and catalytic turnover. The quantity of alpha helices and βsheets produced throughout the simulation time was calculated and displayed in **S-Figure 1**. Analysis revealed a number of secondary structure elements, such as T for the structure's turns and loops, H represents α helix in the SS motif, B represents the isolated H-bonding β bridge between β strands, and E represents the β-strands. Meanwhile, the protein backbone promotes a helical structure, G denotes a particular kind of helical Sec-Structure, Pi–Pi denotes a rare kind of helical structure, and C displays coils. The secondary structures of Hit-1, Hit-2, Hit-3, and Control are shown in **S-1 Figure (A, B, C, & D)**.

3.12. Salt Bridges Assessment

Salt bridges, the strongest non-covalent connections in nature, are essential for molecular recognition, protein folding, and protein-protein interactions (Adhav & Saikrishnan, 2023). Salt-bridge interactions involve two different kinds of side chains of amino acids: Glu or Glu in the case of a positively charged ligand, and Arg or Lys when the ligand is negatively charged (Zhang, 2025). In **Figure 9,** histograms for complexes were analyzed based on their binding score as Control (A), Hit-1 (B), Hit-2 (C), and Hit-3 (D). All the systems showed stability during the simulation period, with a stable electrostatic interaction maintained in the consistency range of salt bridge formation between the ranges of 15-21, while B and C exhibited the highest counts that reflected a strong stabilization and higher flexibility, respectively, overall indicating the molecules remain stable and reflecting with ligand interaction stability.

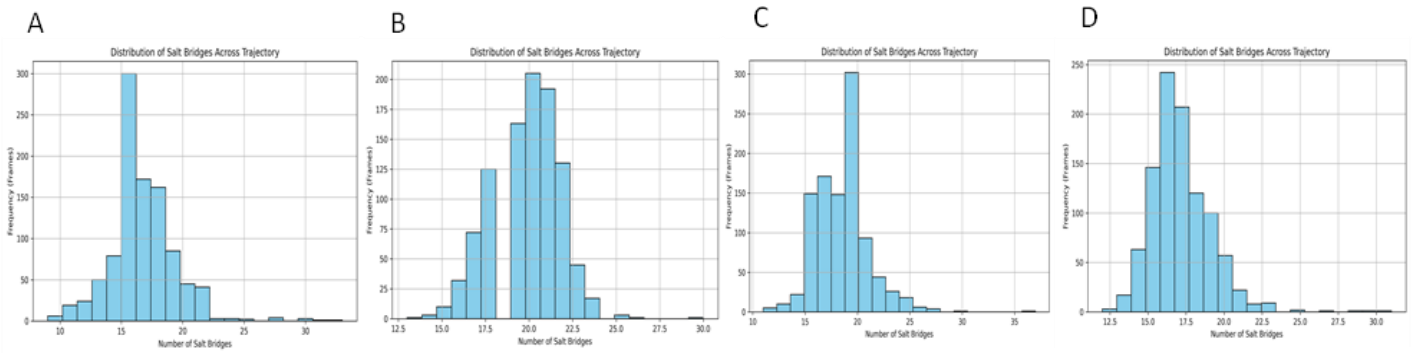


Figure 9. A salt bridge count histogram is illustrated for a selected molecule, while (A) indicates Control, (B) Hit-1, (C) Hit-2, (D) Hit-3, respectively.

4. DISCUSSION

It typically requires a significant amount of time and effort to utilize experimental research to investigate the molecular basis of disease at the structural level. There is a need for in silico methods that can more precisely and effectively identify potential inhibitors, as experimental methods have several limitations (Chang et al., 2023). Several algorithms from several sequence-based and structure-based prediction approaches are combined to provide a useful tool that produces precise and reliable predictions (Micsonai et al., 2022). Depending on whether tumor cells express receptors or not, it is typically divided into five major subtypes: triple negative breast cancer (TNBC), luminal A (LumA), luminal B (LumB), HER2-overexpressing (also known as HER2-enriched or HER2+), and basal-like and normal-like (also known as normal breast tissue-like) (Cai et al., 2024). Hormone receptors (HRs), such as progesterone receptor (PR) and/or estrogen receptor (ER), are present in the luminal A, B, and HER2+ subtypes. HER2 overexpression is also observed in HER2+ and luminal B subtypes. Luminal A and B breast cancer subtypes account for the majority (60–70%) of breast tumors. The HER2 gene is amplified or activated in the HER2+ breast cancer subtype, which accounts for 10–15% of invasive breast cancers. This leads to the overexpression of the HER2 receptor on the surface of breast cancer cells. Up to 90% of cancer fatalities are caused by medication resistance. Chemotherapeutic medication efficacy is hampered by multidrug resistance (MDR), which frequently results in metastases and recurrence. About half of the individuals with medication resistance had both innate and acquired resistance (Eslami et al., 2024). Genetic alterations, the growth of pre-existing insensitive subpopulations like cancer stem cells, and the activation of natural defenses against harmful foreign substances can all result in intrinsic resistance, which happens before therapy (Solary & Laplane, 2020).

A comprehensive in silico study using computational techniques and machine learning (ML) models was conducted to find possible inhibitors against BC from phytochemical compounds. The following models were employed for testing and training ratio of 80:20, with their accuracy scores such as Random Forest (RF), XGBoost, LightGBM, and Artificial Neural Network, and obtained the following training scores as 94%, 94%, 98%, and 86% while the testing values were noted as 91%, 93%, 94%, and 87% respectively. The following models were interpreted and employed a high-throughput screening molecular dicking approach and identified the 3 lead complexes as Hit-1, Hit-2, Hit-3, and as well as screened the Control reported compound against the target protein having an ID 7RKE with the binding affinity of -10 kcal/mol, -9.4 kcal/mol, -9.2 kcal/mol, and -6 kcal/mol and examined within the active pocket binding. All of the identified compounds, apart from the Control molecule, met the requirements to be classified as drug-like, since they had acceptable pharmacokinetic parameters and no discernible change in the binding mode or interactions during the simulation period. The hit compounds were subject to the simulation studies under a 150 ns time frame, and an MMPBSA/GBSA was calculated for the selected complexes. A trajectory data analysis revealed good binding conformations and structural alterations inside the active site during ligand binding. Additionally, PCA and FEL were used to evaluate the data regarding conformational changes in protein-ligand complexes. Following that, the DCCM and RDF studies were carried out to determine whether the protein's structure is unaffected by the novel chemicals.

The study conducted is aligned with the Investigation of Effective Molecular Dynamics-derived Properties on Drug Solubility via Machine Learning (Sodaei et al., 2025) and the development of ML models for solubility prediction, such as Random Forest, Extra Trees, XGBoost, and Gradient Boosting. The purpose of this study is to use machine learning (ML) techniques to statistically investigate the significance of ten MD-derived features and LogP, one of the most important experimental parameters, on medication aqueous solubility. To achieve this, a dataset of 199 compounds from various classes was assembled from the literature, examined using MD simulation, and pertinent characteristics were identified and chosen as features. Furthermore, this study incorporated and considered matching octanol-water partition coefficients (Log P) from previous investigations. Properties that have the biggest impact on solubility were found by explicit analysis. These results indicate that seven characteristics (Log P, SASA, Columbic t, DGSolv, RMSD, AVG shell, and LJ have high predictive power of solubility. The best predictor algorithm that was used in the test set is Gradient Boosting, which gave a prediction R2 = 0.87 and RMSE = 0.537. The study conducted proves the idea that the accuracy and effectiveness of aqueous solubility prediction during drug development can be enhanced with the help of machine learning methods combined with MD simulations.

As shown in our study, the safety and efficacy of the lead compounds identified in in silico studies need comprehensive validation studies to determine their feasibility. Although the in silico analysis has huge potential, there are numerous deficiencies in this approach. Two important attributes of computational correctness are the quality of the structure of the protein and the parameters used in simulating the protein. False positive results are more likely to occur when solvent effects and protein flexibility are ignored in docking studies. The lack of experimental data demands multidisciplinary collaboration to validate antibiotic potency. Future research should mostly focus on drug feature optimization and experimental validation. Combining experimental and computational data might speed up the development of novel therapies for illness-related conditions. In conclusion, our findings contribute to the development of drugs by broadening the variety of treatments accessible for illnesses linked to degradation, even though additional study is needed to address antibiotic resistance. Ultimately, when choosing a large-scale dataset with the available parameter values, the limitations of the conducted study should be taken into account. Virtual screening should capture a complete structure rather than a partial structure, and the simulation time frame should be extended in conjunction with an in vitro screening.

5. CONCLUSION

In the conducted study, several machine learning and deep learning models, i.e., Random Forest, XGBoost, LightGBM, and Artificial Neural Network, with good predicting outcome of models all the models were more than 80% of accuracy in training and testing mode, and implemented for predicting the solubility of phytochemical compounds against BC, then compounds were screened against the target protein 7RKE via protein-ligand docking approach and obtained the three top hits as Hit-1, Hit-2, Hit-3 and Control molecule with good binding score in the active pocket and showed promising results against BC the findings provide a variety of novel inhibitors by targeting the protein domain to slow down the activity of enzyme. Future studies should include experimental validation for the confirmation of computational outputs.

6. REFERENCES

Abourehab, M. A. S., Shawky, A. M., Venkatesan, K., Yasmin, S., Alsubaiyel, A. M., & AboRas, K. M. (2022). Efficiency development of surface tension for different ionic liquids through novel model of machine learning technique: application of in-thermal engineering. *Journal of Molecular Liquids*, *367*, 120391.

Adhav, V. A., & Saikrishnan, K. (2023). The realm of unconventional noncovalent interactions in proteins: their significance in structure and function. *Acs Omega*, *8*(25), 22268–22284.

ALAMUKII, N. A. (2023). *BARRIERS TO EARLY DIAGNOSIS, TUMOUR NECROSIS FACTOR AND RECEPTOR GENETIC VARIANTS AS POSSIBLE PREDICTORS FOR BREAST CANCER AMONG NIGERIAN WOMEN*.

Aldakheel, F. M., & Alduraywish, S. A. (2025). Discovery of novel DdlA inhibitors in multidrug-resistant Pseudomonas aeruginosa using virtual screening, molecular docking, and dynamics simulations. *Scientific Reports*, *15*(1), 15290.

An, F., Sayed, B. T., Parra, R. M. R., Hamad, M. H., Sivaraman, R., Foumani, Z. Z., Rushchitc, A. A., El-Maghawry, E., Alzhrani, R. M., & Alshehri, S. (2022). Machine learning model for prediction of drug

solubility in supercritical solvent: Modeling and experimental validation. *Journal of Molecular Liquids*, *363*, 119901.

Arao, Y., & Korach, K. S. (2021). The physiological role of estrogen receptor functional domains. *Essays in Biochemistry*, *65*(6), 867–875.

Arzanova, E., & Mayrovitz, H. N. (2022). The epidemiology of breast cancer. *Exon Publications*, 1–19.

Belluti, S., Imbriano, C., & Casarini, L. (2023). Nuclear estrogen receptors in prostate cancer: from genes to function. *Cancers*, *15*(18), 4653.

Bizuayehu, H. M., Ahmed, K. Y., Kibret, G. D., Dadi, A. F., Belachew, S. A., Bagade, T., Tegegne, T. K., Venchiarutti, R. L., Kibret, K. T., & Hailegebireal, A. H. (2024). Global disparities of cancer and its projected burden in 2050. *JAMA Network Open*, *7*(11), e2443198–e2443198.

Burley, S. K., Bhikadiya, C., Bi, C., Bittrich, S., Chen, L., Crichlow, G. V, Christie, C. H., Dalenberg, K., Di Costanzo, L., & Duarte, J. M. (2021). RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Research*, *49*(D1), D437–D451.

Cai, A., Chen, Y., Wang, L. S., Cusick, J. K., & Shi, Y. (2024). Depicting biomarkers for HER2-inhibitor resistance: Implication for therapy in HER2-positive breast cancer. *Cancers*, *16*(15), 2635.

Chang, Y., Hawkins, B. A., Du, J. J., Groundwater, P. W., Hibbs, D. E., & Lai, F. (2023). A guide to in silico drug design. *Pharmaceutics*, *15*(1), 49.

Christgen, M., Cserni, G., Floris, G., Marchio, C., Djerroudi, L., Kreipe, H., Derksen, P. W. B., & Vincent-Salomon, A. (2021). Lobular breast cancer: histomorphology and different concepts of a special spectrum of tumors. *Cancers*, *13*(15), 3695.

Dohnalová, H., Seifert, M., Matouskova, E., Klein, M., Papini, F. S., Lipfert, J., Dulin, D., & Lankas, F. (2024). Temperature-dependent twist of double-stranded RNA probed by magnetic tweezer experiments and molecular dynamics simulations. *The Journal of Physical Chemistry B*, *128*(3), 664–675.

Eslami, M., Memarsadeghi, O., Davarpanah, A., Arti, A., Nayernia, K., & Behnam, B. (2024). Overcoming chemotherapy resistance in metastatic cancer: a comprehensive review. *Biomedicines*, *12*(1), 183.

Farooq, A., Mateen, R. M., Ali, M., Javed, M., Dera, A. A., Asimov, A., Ali, S. K., Jaber, F., Bibi, S., & Bahadur, A. (2025). Detection of probable phytochemical inhibitors targeting kallikrein related peptidase 7 (KLK7) in ovarian cancer through molecular dynamics and virtual screening approaches. *Scientific Reports*, *15*(1), 34749.

Gogoi, B., Chowdhury, P., Goswami, N., Gogoi, N., Naiya, T., Chetia, P., Mahanta, S., Chetia, D., Tanti, B., & Borah, P. (2021). Identification of potential plant-based inhibitor against viral proteases of SARS-CoV-2 through molecular docking, MM-PBSA binding energy calculations and molecular dynamics simulation. *Molecular Diversity*, *25*(3), 1963–1977.

Grewal, S., Deswal, G., Grewal, A. S., & Guarve, K. (2025). Molecular dynamics simulations: Insights into protein and protein ligand interactions. In *Advances in Pharmacology* (Vol. 103, pp. 139–162). Elsevier.

Gurung, A. B., Ali, M. A., Lee, J., Farah, M. A., & Al-Anazi, K. M. (2021). An updated review of computer-aided drug design and its application to COVID-19. *BioMed Research International*, *2021*(1), 8853056.

Hajihosseinlou, M., Maghsoudi, A., & Ghezelbash, R. (2023). A novel scheme for mapping of MVT-type Pb–Zn prospectivity: LightGBM, a highly efficient gradient boosting decision tree machine learning algorithm. *Natural Resources Research*, *32*(6), 2417–2438.

He, X., Liu, S., Lee, T.-S., Ji, B., Man, V. H., York, D. M., & Wang, J. (2020). Fast, accurate, and reliable protocols for routine calculations of protein–ligand binding affinities in drug design projects using AMBER GPU-TI with ff14SB/GAFF. *ACS Omega*, *5*(9), 4611–4619.

Hu, X., Chai, X., Wang, X., Duan, M., Pang, J., Fu, W., Li, D., & Hou, T. (2020). Advances in the computational development of androgen receptor antagonists. *Drug Discovery Today*, *25*(8), 1453–1461.

Islam, K., Ramchiary, N., & Kumar, A. (2023). Databases Relevant to Phytochemicals and Genes That Govern Biosynthesis of the Phytochemicals. In *Phytochemical Genomics: Plant Metabolomics and Medicinal Plant Genomics* (pp. 361–377). Springer.

Ji, H., Hu, C., Yang, X., Liu, Y., Ji, G., Ge, S., Wang, X., & Wang, M. (2023). Lymph node metastasis in cancer progression: molecular mechanisms, clinical significance and therapeutic interventions. *Signal*

Transduction and Targeted Therapy, *8*(1), 367.
Kapoor, K., Thangapandian, S., & Tajkhorshid, E. (2022). Extended-ensemble docking to probe dynamic variation of ligand binding sites during large-scale structural changes of proteins. *Chemical Science*, *13*(14), 4150–4169.
King, E., Aitchison, E., Li, H., & Luo, R. (2021). Recent developments in free energy calculations for drug discovery. *Frontiers in Molecular Biosciences*, *8*, 712085.
Klebe, G. (2025). Protein–ligand interactions as the basis for drug action. In *Drug design: from structure and mode-of-action to rational design concepts* (pp. 39–65). Springer.
Lamichhane, T. R., & Ghimire, M. P. (2021). Evaluation of SARS-CoV-2 main protease and inhibitor interactions using dihedral angle distributions and radial distribution function. *Heliyon*, *7*(10).
Li, D.-D., Wu, T.-T., Yu, P., Wang, Z.-Z., Xiao, W., Jiang, Y., & Zhao, L.-G. (2020). Molecular dynamics analysis of binding sites of epidermal growth factor receptor kinase inhibitors. *ACS Omega*, *5*(26), 16307–16314.
Matamoros-Recio, A., Mínguez-Toral, M., & Martín-Santamaría, S. (2023). Modeling of Transmembrane Domain and Full-Length TLRs in Membrane Models. In *Toll-Like Receptors: Methods and Protocols* (pp. 3–38). Springer.
Mazurek, A. H. (2024). *A study of selected endocrine disrupting chemicals and their binding to host molecules with molecular modelling.* Institut Polytechnique de Paris; Medical University of Warsaw.
Micsonai, A., Moussong, E., Wien, F., Boros, E., Vadászi, H., Murvai, N., Lee, Y.-H., Molnár, T., Réfrégiers, M., & Goto, Y. (2022). BeStSel: webserver for secondary structure and fold prediction for protein CD spectroscopy. *Nucleic Acids Research*, *50*(W1), W90–W98.
Mienye, I. D., & Sun, Y. (2022). A survey of ensemble learning: Concepts, algorithms, applications, and prospects. *Ieee Access*, *10*, 99129–99149.
Miziak, P., Baran, M., Błaszczak, E., Przybyszewska-Podstawka, A., Kałafut, J., Smok-Kalwat, J., Dmoszyńska-Graniczka, M., Kiełbus, M., & Stepulak, A. (2023). Estrogen receptor signaling in breast cancer. *Cancers*, *15*(19), 4689.
Moradi, S., Nowroozi, A., Nezhad, M. A., Jalali, P., Khosravi, R., & Shahlaei, M. (2024). A review on description dynamics and conformational changes of proteins using combination of principal component analysis and molecular dynamics simulation. *Computers in Biology and Medicine*, *183*, 109245.
Mosa, F. E. S., Alqahtani, M. A., El-Ghiaty, M. A., Barakat, K., & El-Kadi, A. O. S. (2024). Identifying novel aryl hydrocarbon receptor (AhR) modulators from clinically approved drugs: In silico screening and in vitro validation. *Archives of Biochemistry and Biophysics*, *754*, 109958.
Mutharasappan, N., Ravi Rao, G., Mariadasse, R., Poopandi, S., Mathimaran, A., Dhamodharan, P., Sundarraj, R., Jeyaraj Pandian, C., & Jeyaraman, J. (2020). Experimental and computational methods to determine protein structure and stability. *Frontiers in Protein Structure, Function, and Dynamics*, 23–55.
Niazi, S. K. (2025). Protein catalysis through structural dynamics: a comprehensive analysis of energy conversion in enzymatic systems and its computational limitations. *Pharmaceuticals*, *18*(7), 951.
Palaniappan, M. (2024). Current Therapeutic Opportunities for Estrogen Receptor Mutant Breast Cancer. *Biomedicines*, *12*(12), 2700.
Pantaleão, S. Q., Fernandes, P. O., Gonçalves, J. E., Maltarollo, V. G., & Honorio, K. M. (2022). Recent advances in the prediction of pharmacokinetics properties in drug design studies: a review. *ChemMedChem*, *17*(1), e202100542.
Qamar, R., & Zardari, B. A. (2023). Artificial neural networks: An overview. *Mesopotamian Journal of Computer Science*, *2023*, 124–133.
Rane, N., Choudhary, S. P., & Rane, J. (2024). Ensemble deep learning and machine learning: applications, opportunities, challenges, and future directions. *Studies in Medical and Health Sciences*, *1*(2), 18–41.
Sad, N. (2024). Female breast cancer: an updated review of epidemiology, risk factors and prevention. *Hippokratia*, *28*(4), 135–142.
Sadat, A., & Joye, I. J. (2020). Peak fitting applied to fourier transform infrared and raman spectroscopic analysis of proteins. *Applied Sciences*, *10*(17), 5918.
Sahu, M., & Suryawanshi, H. (2021). Immunotherapy: The future of cancer treatment. *Journal of Oral and Maxillofacial Pathology*, *25*(2), 371.

Salehi, S. M., & Meuwly, M. (2022). Cross-correlated motions in azidolysozyme. *Molecules*, *27*(3), 839.
Sandunil, K., Bennour, Z., Mahmud, H. Ben, & Giwelli, A. (2024). Effects of tuning decision trees in random forest regression on predicting porosity of a hydrocarbon reservoir. A case study: volve oil field, north sea. *Energy Advances*, *3*(9), 2335–2347.
Sardar, H. (2023). Drug like potential of Daidzein using SwissADME prediction: In silico Approaches. *Phytonutrients*, 2–8.
Sasidharan, S., Gosu, V., Tripathi, T., & Saudagar, P. (2023). Molecular Dynamics simulation to study protein conformation and ligand interaction. In *Protein folding dynamics and stability: experimental and computational methods* (pp. 107–127). Springer.
Shukla, R., & Tripathi, T. (2020). Molecular dynamics simulation of protein and protein–ligand complexes. In *Computer-aided drug design* (pp. 133–161). Springer.
Singh, A., & Roghini, S. (2023). Cancer: Unraveling the Complexities of Uncontrolled Growth and Metastasis. *PEXACY International Journal of Pharmaceutical Science*, *2*(8), 59–73.
Sodaei, Z., Ekrami, S., & Hashemianzadeh, S. M. (2025). *Investigation of Effective Molecular Dynamics-derived Properties on Drug Solubility via Machine Learning*.
Solary, E., & Laplane, L. (2020). The role of host environment in cancer evolution. *Evolutionary Applications*, *13*(7), 1756–1770.
Sucharitha, P., Reddy, K. R., Satyanarayana, S. V, & Garg, T. (2022). Absorption, distribution, metabolism, excretion, and toxicity assessment of drugs using computational tools. In *Computational approaches for novel therapeutic and diagnostic designing to mitigate SARS-CoV-2 infection* (pp. 335–355). Elsevier.
Toumba, M., Kythreotis, A., Panayiotou, K., & Skordis, N. (2024). Estrogen receptor signaling and targets: Bones, breasts and brain. *Molecular Medicine Reports*, *30*(2), 144.
Tyagi, A. K., & Chahal, P. (2020). Artificial intelligence and machine learning algorithms. In *Challenges and applications for implementing machine learning in computer vision* (pp. 188–219). IGI Global Scientific Publishing.
Velázquez, C., K, D. L., Esteban-Cardeñosa, E. M., Avila Cobos, F., Lastra, E., Abella, L. E., de la Cruz, V., Lobatón, C. D., Claes, K. B., & Durán, M. (2020). Germline genetic findings which may impact therapeutic decisions in families with a presumed predisposition for hereditary breast and ovarian cancer. *Cancers*, *12*(8), 2151.
Vora, L. K., Gholap, A. D., Jetha, K., Thakur, R. R. S., Solanki, H. K., & Chavda, V. P. (2023). Artificial intelligence in pharmaceutical technology and drug delivery design. *Pharmaceutics*, *15*(7), 1916.
Wu, C., Dong, S., Huang, R., & Chen, X. (2023). Cancer-associated adipocytes and breast cancer: intertwining in the tumor microenvironment and challenges for cancer therapy. *Cancers*, *15*(3), 726.
Xie, P., An, R., Yu, S., He, J., & Zhang, H. (2021). A novel immune subtype classification of ER-positive, PR-negative and HER2-negative breast cancer based on the genomic and transcriptomic landscape. *Journal of Translational Medicine*, *19*(1), 398.
Yu, K., Huang, Z.-Y., Xu, X.-L., Li, J., Fu, X.-W., & Deng, S.-L. (2022). Estrogen receptor function: impact on the human endometrium. *Frontiers in Endocrinology*, *13*, 827724.
Zhang, Y. (2025). *THERMODYNAMIC AND STRUCTURAL PROPERTIES OF ARGININE RESIDUES IN THE PROTEIN INTERIOR*. Johns Hopkins University.
Zhao, P., & Malik, S. (2022). The phosphorylation to acetylation/methylation cascade in transcriptional regulation: how kinases regulate transcriptional activities of DNA/histone-modifying enzymes. *Cell & Bioscience*, *12*(1), 83.
Zhuo, C., Zeng, C., Yang, R., Liu, H., & Zhao, Y. (2023). RPflex: A coarse-grained network model for RNA pocket flexibility study. *International Journal of Molecular Sciences*, *24*(6), 5497.