

Machine learning-based 3D-QSAR models for predicting the estrogen receptor-binding activity of small molecules

BR. Bharath^{*}, Sreerupa Mitra, Nadeem Khan

Jai Research Foundation, Valvada, Vapi, Gujarat 396105, India

ARTICLE INFO

Keywords:

Estrogen receptor alpha
2D-QSAR
3D-QSAR
VEGA

ABSTRACT

Estrogen receptor alpha (ER α) belongs to the steroid receptor superfamily and acts as a ligand-activated transcription factor. Structurally, ER α comprises six domains labeled A through F. The DNA-binding domain (DBD) facilitates specific interactions with estrogen response elements, while the ligand-binding domain (LBD) engages with various agonistic and antagonistic hormones. Notably, numerous endocrine-disrupting chemicals (EDCs) exert adverse effects on estrogen signaling by interacting with ER α . Consequently, there is a critical need to evaluate the endocrine disruption potential of new chemical entities (NCEs) by scrutinizing their interactions with multiple pertinent targets. ER α is a pivotal target protein, and leveraging third-party tools for predicting the relative binding affinity (RBA) of small molecules via 2D-QSAR has become common. However, in this study, we advanced beyond conventional methods by developing machine learning-based 3D-QSAR models. These 3D-QSAR models were built using the classification dataset of VEGA V.1.2.0 for the estrogen receptor IRFMN—CERAPP and IRFMN-RBA models. Our investigation demonstrated that the 3D-QSAR models, which employ algorithms such as random forest (RF), support vector machine (SVM), and multilayer perceptron (MLP), outperform the VEGA models in terms of accuracy, sensitivity, and selectivity. Furthermore, the efficacy of these models was corroborated through validation against external datasets. Notably, the 3D-QSAR models exhibit superior accuracy and sensitivity compared to the VEGA model, thereby offering a promising approach for assessing the endocrine disruption potential of novel chemical entities.

1. Introduction

Estrogen receptors (ERs) are nuclear hormone receptors that are responsible for regulating gene expression and influencing cellular proliferation and differentiation in eukaryotic organisms. Certain endocrine-disrupting chemicals (EDCs) can activate these receptors, disrupting normal estrogen signaling [1]. ER α and estrogen receptor beta (ER β) are two distinct types of estrogen receptors that share high similarity in their DNA binding domains but differ in other regions, leading to selective binding by EDCs. ER α binders are better characterized than ER β binders, with EDCs causing deregulation of response elements and resulting in neurological, developmental, and reproductive toxicity [2].

The NCEs could potentially be EDCs, as they may interact with the ligand binding domain of ER α [3,4]. Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), a European chemical regulator, identifies EDCs and others concerning chemicals requiring stringent permission and supervision. Evaluating the estrogen receptor

binding potential of NCEs conventionally is costly and time-consuming.

High-throughput screening (HTS) offers an alternative by rapidly testing numerous molecules at lower costs. HTS has generated vast amounts of *in vitro* data accessible in bioassay databases, enabling the development of QSAR models for predicting EDC potential before synthesis [5,7]. With the rise of HTS data, virtual screening techniques have emerged as an even more advanced and reliable alternative. The Vega algorithm has developed QSAR binary classifier models using various data mining techniques, including support vector machines, fuzzy logic, decision forests, neural networks, and classification trees [6]. These models assess multiple endpoints to robustly evaluate the effects of EDCs, including transcriptional activity and binding interactions [5].

The Collaborative Estrogen Receptor Activity Prediction Project (CERAPP), conducted by VEGA, undertook a significant modeling effort demonstrating the efficacy of predictive computational models trained on high-throughput screening (HTS) data to assess thousands of chemicals for estrogen receptor (ER)-mediated transcriptional activity. CERAPP integrates various models developed by 17 groups from the U.S.

^{*} Corresponding author.

E-mail address: bharath.rudresh@jrffonline.com (BR. Bharath).

and Europe, employing QSAR models and docking approaches primarily using a common training set provided by the U.S. EPA [8]. This effort resulted in the creation of 40 categorical and 8 continuous models for ER binding, agonist, and antagonist activity. The reliability and robustness of these models were evaluated using a dataset of 7522 compounds selected from the literature, and a consensus was established by weighing the models based on their accuracy [8].

Additionally, VEGA developed the relative binding affinity (RBA) model to predict the relative binding affinity of EDCs. This model utilized a dataset from the METI database comprising experimentally determined values of human ER-alpha activity, expressed as a percentage of activity using 17-estradiol as a reference [5,8]. The dataset represented a diverse range of compounds, including natural and synthetic steroids, drugs, and chemical contaminants. Both the CERAPP and RBA models employ a simplistic approach by considering only the 2D configuration of molecules, disregarding the 3D conformation and steric configuration [5]. However, the inclusion of 3D descriptors is essential for enhancing the reliability and robustness of QSAR models [9]. The 3D-QSAR method involves a matrix corresponding to molecular fields defined by the 3D properties of the system, allowing for better characterization of nonbonded interactions between the ligand and the receptor [10–12].

Earlier studies have reported the application of 3D-QSAR models as a tool for virtual screening of molecules and hit-to lead optimization. In some recent studies 3D-QSAR models have been developed to predict the activity of a molecule on ER α as antagonists using 3D-pharmacophore based approaches for 97 known ER α binders [13], CoMFA and CoMSIA methods-based 3D-QSAR regression models for 81 compounds [14]. However, most of these models were built using a small dataset with less structural diversity.

To address this limitation, the present study used the CERAPP and RBA datasets from VEGA V1.2.0 to develop a 3D-QSAR model aimed at enhancing predictability. Machine learning algorithms such as k-nearest neighbors (kNN), SVM, RF, and MLP were employed to develop multiple models with high reliability and robustness.

2. Materials and methods

2.1. Dataset collection and cleaning

The datasets used for developing the 3D-QSAR models were obtained through VEGA V1.2.0. The dataset corresponding to the estrogen Istituto di Ricerche Farmacologiche Mario Negri - Collaborative Estrogen Receptor Activity Prediction Project (IRFMN—CERAPP) model used to predict ER α transcriptional activity included 1529 highly diverse molecules. Among these, 89 molecules were labeled ‘active’, while the remaining 1440 were labeled ‘inactive’ [8]. Similarly, for predicting ER α receptor binding activity, the dataset associated with the estrogen Istituto di Ricerche Farmacologiche Mario Negri – Relative Binding Affinity (IRFMN-RBA) model was used, consisting of 806 highly diverse molecules. Within this dataset, 290 molecules were labeled ‘active’, and the remaining 516 were labeled ‘inactive’ [5]. Notably, six molecules containing tin (Sn) were initially included. However, due to tool incompatibility, these tin-containing molecules were excluded from the present study. The final dataset size for predicting ER α receptor binding activity was subsequently adjusted to 800 molecules (Fig. 1).

There are many approaches for evaluating the reliability and robustness of the QSAR model to predict the desired endpoint. Among them, validating QSAR models using external datasets is a powerful approach. In the present study, two different external datasets were used to evaluate the predictability of the models developed for the IRFMN—CERAPP and IRFMN-RBA datasets. The external dataset for IRFMN—CERAPP was derived from the VEGA evaluation set for IRFMN—CERAPP. Overall, 163 molecules tagged as agonists in the IRFMN—CERAPP evaluation set were considered. Furthermore, the external dataset for IRFMN-RBA was retrieved from the PDB, and the 52

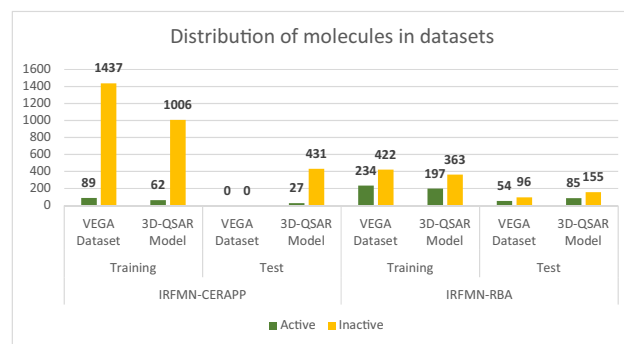


Fig. 1. Distribution of molecules in IRFMN—CERAPP and IRFMN-RBA dataset with respect to training and test sets. The distribution of active and inactive in the dataset is depicted by differentially colored bars.

co-crystals in complex with human estrogen receptor alpha (hER α) were considered positive for receptor binding affinity. The external dataset was screened for two endpoints available in VEGA v1.2.0, namely, ER-mediated effects and ER-related binding affinity [15]. In parallel, the same external dataset was subjected to prediction using 3D-QSAR models developed during the study (Fig. 2).

2.2. Dataset partition

The data partitioning method involves dividing the dataset into subsets for training and testing, typically described by a split ratio. The commonly employed ratio is 80:20, indicating that 80 % of the data are allocated to the training dataset, while the remaining 20 % form the testing dataset. Although other ratios, such as 70:30, 60:40, and even 50:50, are also utilized in practice, there is no definitive guidance on selecting the optimal ratio. Instead, the choice of dataset depends on the relevance and diversity of the dataset [16]. In this study, dataset partitioning was optimized using ratios of 80:20, 70:30, and 60:40, with activity classified as either 1 (active) or 0 (inactive). These ratios were chosen based on the Pareto principle, or the 80:20 rule, which is commonly used by practitioners, though its suitability may depend on the specific dataset.

2.3. Loop modeling of human ER α (hER α)

Various methods, including computational molecular field analysis (CoMFA), comparative molecular indices analysis (CoMSIA), pharmacophore generation, and free-energy binding analysis, are employed in conducting 3D-QSAR studies [17]. Hence, determining the structure of the target protein is imperative. To this end, the human estrogen receptor alpha (hER α) protein sequence, with a length of 595 amino acids (UniProt ID: P03372), was queried against the Protein Data Bank (PDB), yielding 433 available entries. Among these, numerous entries

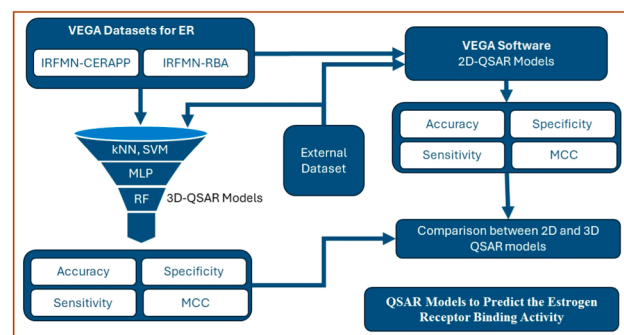


Fig. 2. Illustration of workflow for the development and validation of QSAR models.

contained structures for partial sequences and various inhibitors as cocrystals. Given the focus on the ligand-binding domain in this study, a curated entry containing raloxifene as a cocrystal (PDB ID: 1ERR) was selected. Raloxifene is a classic example of a selective estrogen receptor modulator (SERM) that has both agonistic and antagonistic effects on ER α . Specifically, it has antagonistic effects on the breast and uterus but has agonistic effects on bone [18]. Although the curated structure lacked certain loops, necessitating their modeling, the conformations were selected from a FREAD database of experimentally determined protein fragments. This modeling process was executed utilizing the loop modeling module available in Cresset Flare Pro Plus software [19].

2.4. Molecular docking

The structure of hER α was prepared to obtain the bioactive conformation via postloop modeling. This could appear as a single representation, yet it is more plausible that this representation is an ensemble of closely related structures that showcase the concerted movements of both the protein and ligand upon binding. During preparation, hydrogens corresponding to pH 7.0 were incorporated, considering the appropriate ionization states for both acidic and basic amino acid residues [20]. Residue gaps were filled, and atoms from residues with incomplete backbone atoms were excised. The active site size was established at 6.00 Å, and the prepared protein was duplicated for molecular docking studies. The binding site of the cocrystal of raloxifene served as the active site, and a cocrystal-centric grid box was generated. To validate the docking parameters, low-energy conformations of raloxifene were docked into the active pocket of hER α employing a highly accurate albeit slow method, employing the extra precision quality available in the Cresset Flare module [21]. Simultaneously, molecules from the IRFMN—CERAPP and IRFMN-RBA datasets were also docked to hER α . A maximum of 10 docked conformations were generated, and the final best docked conformation was selected based on the binding energy (Δ_G value).

2.5. Confirmation hunt and ligand alignment

Ligand alignment is a critical process that significantly influences the quality of the 3D-QSAR model. In this study, ligand-based alignment was employed to align the ligands in the IRFMN—CERAPP dataset using the cocrystal conformation of raloxifene as the reference molecule. The alignment process was performed using a Cresset Flare, with the eXtended Electron Distribution (XED) force field employed for field point calculations. To ensure accuracy in conformer selection, the Very Accurate and Slow calculation method was used during the conformation search, aiming to obtain precise conformations conducive to generating a robust 3D-QSAR model. Similarly, conformation hunting and alignment were performed for the molecules in the IRFMN-RBA dataset using 8 cocrystal molecules found in complex with different entries for estrogen receptors in the PDB.

2.6. QSAR modeling

Classification or categorical QSAR models, such as kNN, RF, SVM, and MLP, were constructed to evaluate the potential of small molecules to inhibit hER α . These models incorporated 3D descriptors, including electrostatic and volume properties, to enhance model development. The efficacy of the models was assessed through metrics such as accuracy, sensitivity, specificity, and Matthew's correlation coefficient (MCC). Additionally, the predictive capability of the developed models was evaluated using an external dataset for validation purposes.

3. Results and discussion

3.1. Dataset quality and optimization

The quality and partitioning of the dataset are crucial elements in the performance of predictive models, particularly in the context of imbalanced datasets, as they influence the sensitivity, accuracy, and robustness of the resulting models. In the present study, two distinct datasets, IRFMN—CERAPP and IRFMN-RBA, were analyzed for activity distributions and partition strategies.

The IRFMN—CERAPP dataset, consisting of 1526 molecules, was heavily imbalanced, with 1437 inactive and only 89 active molecules. Such an imbalanced distribution is a known challenge in machine learning and can lead to poor sensitivity in predictive models, as the model may become biased towards predicting the majority class (inactive molecules). To mitigate this, the dataset was partitioned into training and test sets at a 70:30 ratio, aiming to balance training data exposure while maintaining enough testing data for validation. In contrast, the IRFMN-RBA dataset, consisting of 806 molecules with a more balanced distribution (288 active and 518 inactive molecules), was split with three partition ratios—80:20, 70:30, and 60:40. This partitioning allowed for an empirical comparison of model performance across different training data sizes, shedding light on the impact of dataset size and balance on model accuracy and sensitivity.

As illustrated in Fig. 3, the models developed using partition ratios of 80:20 (for IRFMN-RBA) yielded the best performance in terms of both sensitivity and accuracy. For example, the RF, SVM, and MLP models achieved sensitivity values of 0.93, 0.94, and 0.95, respectively, with accuracy of 0.97, 0.96, and 0.96 (Table 1). These results underscore the importance of selecting an optimal partition ratio, as models trained on an 80:20 partition achieved the highest balance between model complexity and predictive performance.

3.2. Loop modeling and molecular docking

The process of loop modeling and molecular docking represents a crucial step in understanding the structural elements critical for ligand binding. In this study, the crystal structure of hER α (PDB ID: 1ERR) revealed missing loops between residues Thr460 and Leu469 and between Lys529 and Val534 (Fig. 4A). These missing loop regions, located near the ligand binding site, were modeled to examine their influence on the ligand binding site geometry. The loop modeling resulted in a significant increase in the surface area of the ligand binding site—from

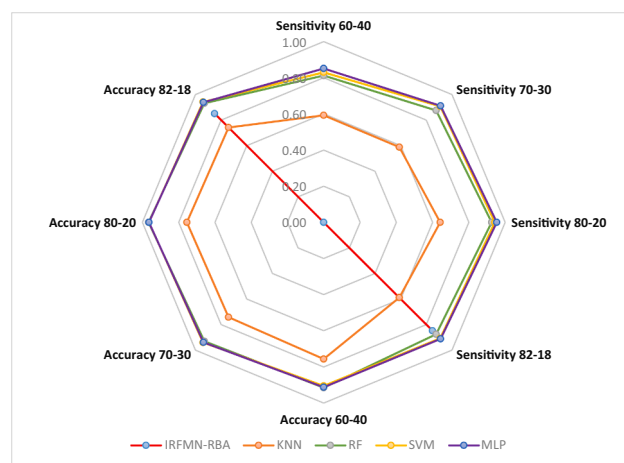


Fig. 3. The plot illustrates the differences between sensitivity and accuracy of different 3D QSAR models with different partition ratios such as 80:20, 70:30 and 60:40. The sensitivity and accuracy of VEGA model with the 82:12 partition ratio is also compared with the 3D-QSAR models.

Table 1

The differences between sensitivity and accuracy of different 3D QSAR models with different partition ratios such as 80:20, 70:30 and 60:40. The sensitivity and accuracy of VEGA model with the 82:12 partition ratio is also compared with the 3D-QSAR models.

		IRFMN-RBA	KNN	RF	SVM	MLP
Sensitivity	60–40	NA	0.59	0.81	0.83	0.85
	70–30	NA	0.59	0.88	0.91	0.91
	80–20	NA	0.64	0.93	0.94	0.95
	82–18	0.85	0.59	0.88	0.91	0.91
Accuracy	60–40	NA	0.76	0.92	0.91	0.91
	70–30	NA	0.74	0.93	0.94	0.94
	80–20	NA	0.76	0.97	0.96	0.96
	82–18	0.85	0.74	0.93	0.94	0.94

541.89 Å to 1092.01 Å—demonstrating the structural significance of these loops in ligand binding as shown in Fig. 4B (the region highlighted with blue and the molecular surface representation).

The impact of loop modeling on the ligand binding site was quantitatively evaluated by comparing the surface areas of the ligand binding site before and after loop insertion. As shown in Fig. 4C and D, the total surface area of hER α increased by 669.02 Å², with a noteworthy increase of 550.21 Å² in the ligand binding site itself. This result highlights the importance of loop modeling in accurately defining ligand binding regions, thus improving the reliability of subsequent molecular docking experiments.

The validation of the docking protocol using the cocrystal structure of raloxifene further corroborated the accuracy of the docking approach (Fig. 5A). The root mean square deviation (RMSD) between the cocrystal conformation (Fig. 5B) and the docked pose (Fig. 5C) was found to be 0.93 Å, well within the acceptable threshold of 2 Å [22]. Moreover, the

docked conformation exhibited interactions with the same amino acids (e.g., Asp351, Glu353, Trp383, Arg394) as the cocrystal structure, validating the docking protocol's ability to reproduce experimentally observed binding interactions. Furthermore, molecular docking was performed for 1437 inactive molecules in the IRFMN—CERAPP dataset, and the conformations with the lowest binding energies were identified. The phenobarbital with the lowest binding energy of -6.28 kJ/mol was able to establish an interaction with only Phe404 in the ligand binding site (Fig. 6) and was considered an inactive reference for the conformation search and alignment of molecules in the IRFMN—CERAPP dataset. Since the cocrystals were used as the reference for the IRFMN-RBA dataset, there was no need to perform molecular docking for this dataset.

3.3. Assessment of the predictivity of 3D-QSAR models

The 3D-QSAR models developed in this study (using kNN, RF, SVM, and MLP algorithms) were evaluated in terms of predictive performance, with a focus on accuracy, specificity, sensitivity, and MCC. These metrics offer a comprehensive evaluation of the model's ability to predict both positive and negative instances, especially in the context of imbalanced datasets, where accuracy alone may not fully reflect model performance [23,24].

3.3.1. Model accuracy

Accuracy, a widely used metric, measures the proportion of correct predictions made by a model, including both true positives and true negatives, over the total number of predictions. Although it is a straightforward measure of model performance, accuracy can sometimes be misleading, especially when dealing with imbalanced datasets

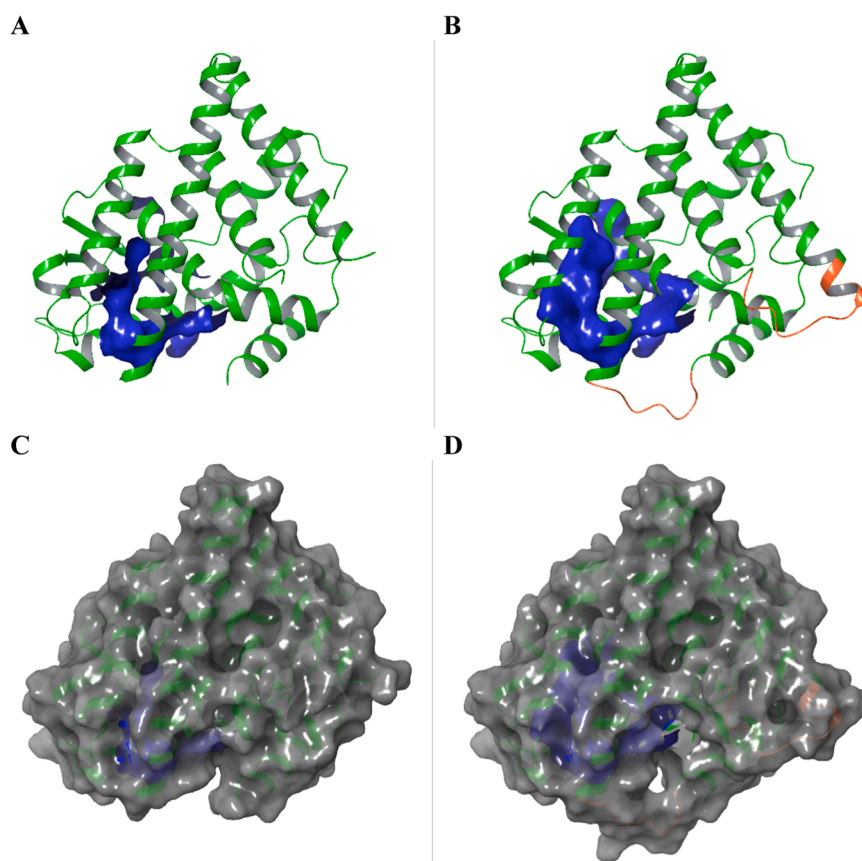
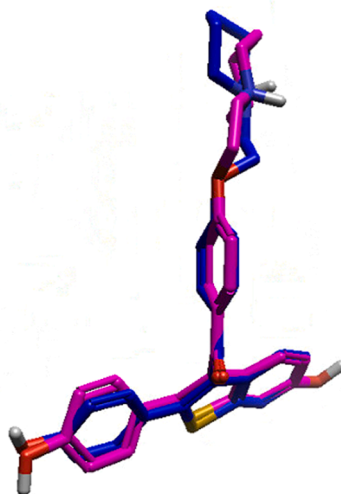
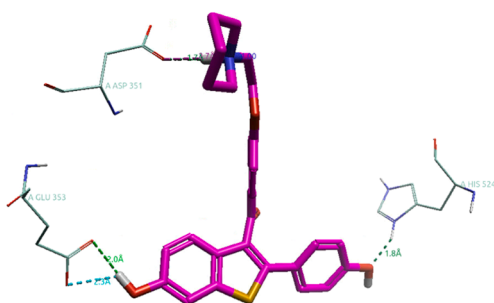


Fig. 4. Modeling of missing loops. **A:** The structure of hER α with missing loops. **B:** The structure of hER α with modelled loops (highlighted in orange color). **C:** The structure of hER α with missing loops with molecular surface representation. **D:** The structure of hER α with modelled loops with molecular surface representation. The blue colored molecular representation corresponds to ligand binding site.

A



B



C

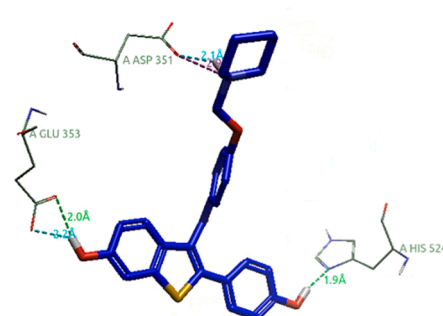
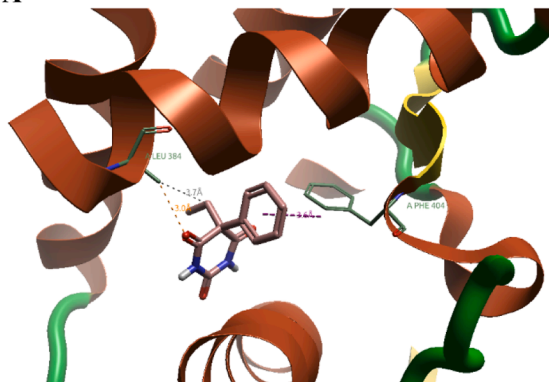


Fig. 5. Validation of docking protocol. **A:** Superimposition of co-crystal (Pink) and docked (Blue) conformations of raloxifene in hER α crystal structure. **B:** The co-crystal conformation of raloxifene showing interaction with binding site amino acids. **C:** The docked conformation of raloxifene showing interaction with binding site amino acids. The dotted lines indicate bonds between raloxifene and amino acids, green: hydrogen bond, blue: weak hydrogen bonds, orange: steric clash, purple: pi-pi stacking and salt bridges.

A



B



Fig. 6. Molecular docking of phenobarbital with hER α . **A:** The three-dimensional representation of docked conformation of phenobarbital showing interaction with binding site amino acids. **B:** The two-dimensional illustration of phenobarbital showing interaction with binding site amino acids. The dotted lines indicate bonds between raloxifene and amino acids, orange: steric clash, purple: pi-pi stacking and salt bridges.

or when different types of errors (false positives and false negatives) carry varying costs. In this study, accuracy was assessed for various models across two datasets: IRFMN—CERAPP and IRFMN-RBA. These datasets were used to predict the potential of small molecules for hER α activity, providing a practical context for evaluating model performance.

The results presented in Figs. 7 and 8, along with Tables 2 and 3, underscore the overall superiority of 3D-QSAR models compared to 2D-

QSAR models in terms of accuracy. This difference highlights the importance of incorporating three-dimensional molecular features, which are better suited to capture the complexity and spatial arrangement of molecules, in predicting the biological activity of small molecules.

3.3.1.1. IRFMN—CERAPP dataset. In the IRFMN—CERAPP dataset, which is imbalanced, the 2D-QSAR model achieved an accuracy of 0.81,

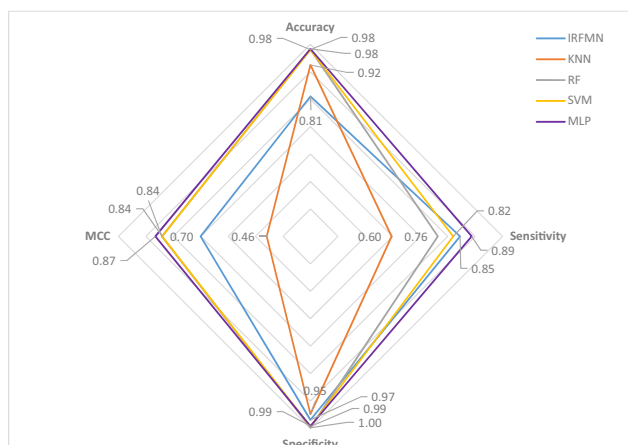


Fig. 7. The plot illustrating the predictive ability and the robustness of different QSAR models developed for IRFMN—CERAPP dataset to predict the potential of small molecule for hER α transcriptional activity.

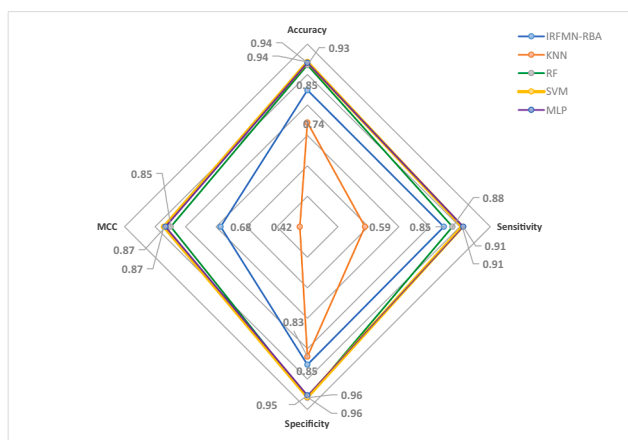


Fig. 8. The plot illustrating the predictive ability and the robustness of different QSAR models developed for IRFMN-RBA dataset to predict the potential of small molecule for hER α binding activity.

Table 2

The robustness of different QSAR models developed for IRFMN—CERAPP dataset to predict the potential of small molecule for hER α transcriptional activity.

	IRFMN	KNN	RF	SVM	MLP
Accuracy	0.81	0.92	0.98	0.98	0.98
Sensitivity	0.85	0.60	0.76	0.82	0.89
Specificity	0.97	0.95	1.00	0.99	0.99
MCC	0.70	0.46	0.84	0.84	0.87

Table 3

The robustness of different QSAR models developed for IRFMN-RBA dataset to predict the potential of small molecule for hER α binding activity.

	IRFMN-RBA	KNN	RF	SVM	MLP
Accuracy	0.85	0.74	0.93	0.94	0.94
Sensitivity	0.85	0.59	0.88	0.91	0.91
Specificity	0.85	0.83	0.96	0.96	0.95
MCC	0.68	0.42	0.85	0.87	0.87

reflecting its ability to make correct predictions for both active and inactive molecules. However, 3D-QSAR models exhibited significantly better performance, with all 3D-QSAR models surpassing 0.90 in

accuracy (Table 2). This result suggests that 3D-QSAR models, which consider the spatial features and the three-dimensional structure of molecules, can better capture the intricate patterns related to molecular activity than the 2D-QSAR model.

The KNN model, a 3D-QSAR technique, achieved an accuracy of 0.92, which is considerably higher than the IRFMN—CERAPP 2D-QSAR model. This improvement can be attributed to the fact that KNN calculates the distance between molecules in a multi-dimensional feature space, which enables it to better account for the complex relationships between molecular structure and biological activity. However, the improvement in accuracy by KNN over 2D-QSAR is modest compared to other models.

The RF, SVM, and MLP models, all based on more sophisticated learning techniques, demonstrated even better accuracy rates, all achieving an impressive accuracy of 0.98 (Table 2). These models, which use decision boundaries or non-linear transformations to classify molecules, leverage advanced statistical techniques and optimization algorithms that further refine the classification, especially when dealing with the high-dimensional data typical of 3D-QSAR models.

3.3.1.2. IRFMN-RBA dataset. In the IRFMN-RBA dataset, which is more balanced than IRFMN—CERAPP, the 2D-QSAR model performed slightly better than the KNN 3D-QSAR model (accuracy of 0.85 vs. 0.74). This is likely because the 2D-QSAR model, which relies on simpler molecular descriptors, can perform adequately when the dataset is balanced, as the class distribution does not heavily bias the model.

In contrast, the 3D-QSAR models, excluding KNN, outperformed the 2D-QSAR model by a significant margin. Random Forest (RF), Support Vector Machine (SVM), and Multi-Layer Perceptron (MLP) achieved accuracies of 0.93, 0.94, and 0.94, respectively (Table 3). These models have superior performance due to their advanced techniques in learning and decision-making, which are better suited to handling the molecular complexity captured by 3D-QSAR features.

3.3.2. Specificity of the models

Specificity, as a key metric in model evaluation, measures the proportion of true negatives (inactive molecules) that are correctly identified by the predictive model. It plays a crucial role in distinguishing between inactive compounds and ensuring that the model does not falsely classify inactive molecules as active. In predictive modeling, particularly for datasets where the distribution of active and inactive molecules is imbalanced, achieving a high specificity is vital to ensure that the model is not overly biased toward predicting the majority class (inactive molecules). This ability to correctly reject false positives—i.e., inactive molecules predicted as active—ensures that the model is robust and accurate in its classification.

In the context of this study, specificity was examined across several models, including kNN, RF, SVM, and MLP for both the IRFMN—CERAPP and IRFMN-RBA datasets. The analysis revealed critical insights into how these models performed in terms of specificity and what that means for their robustness and practical application.

3.3.2.1. kNN model performance. As described in Figs. 7 and 8, the kNN algorithm demonstrated lower specificity values compared to the IRFMN—CERAPP and IRFMN-RBA 2D-QSAR models, with specificity scores of 0.95 and 0.83 for IRFMN—CERAPP and IRFMN-RBA datasets, respectively (Tables 2 and 3). This result suggests that the kNN model struggled to differentiate inactive molecules from active ones, as seen in the higher number of false positives (inactive molecules incorrectly classified as active). The kNN algorithm relies on calculating the distance between feature vectors in a high-dimensional space, but it does not inherently account for model fitting or optimization of decision boundaries. As a result, kNN tends to assign a new data point to the class of its nearest neighbors without distinguishing between class distributions effectively. This lack of explicit boundary construction means that

kNN can misclassify inactive molecules more easily, leading to lower specificity.

3.3.2.2. RF, SVM, and MLP models. In contrast, the RF, SVM, and MLP models demonstrated significantly higher specificity across both datasets. These models employ more sophisticated strategies that optimize decision boundaries and leverage fitting techniques that consider the entire feature space, allowing them to better handle the complexity of molecular activity prediction.

The RF model showed near-perfect specificity of 1.00 for the IRFMN—CERAPP dataset and 0.96 for the IRFMN-RBA dataset, highlighting its ability to correctly identify the majority of inactive molecules (Table 2 and 3). The RF algorithm builds multiple decision trees based on subsets of the data, and its ensemble nature makes it robust to overfit. The random selection of features for each tree ensures that the model does not become biased by any one feature, improving its ability to generalize and correctly classify inactive molecules as negatives.

The SVM model performed similarly well, with specificity values of 0.99 for IRFMN—CERAPP and 0.96 for IRFMN-RBA. SVM works by finding the optimal hyperplane that best separates the classes of data. The model's ability to maximize the margin between the decision boundary and the closest data points (support vectors) makes it less prone to misclassifying inactive molecules as active. This is particularly beneficial in the case of imbalanced datasets, where correctly classifying the minority class (active molecules) is often prioritized. However, the model also maintains a strong ability to correctly classify the majority class (inactive molecules), ensuring high specificity.

The MLP model achieved high specificity values of 0.99 for IRFMN—CERAPP and 0.95 for IRFMN-RBA. MLPs, being a type of artificial neural network, can capture complex non-linear relationships in the data. The architecture of MLPs, with multiple layers of interconnected nodes, allows the model to adapt and learn intricate patterns in the dataset. This leads to better generalization and accurate classification of inactive molecules, reflected in the high specificity values.

3.3.3. Sensitivity of the models

In machine learning, sensitivity (also known as recall or true positive rate) is a vital metric that measures a model's ability to correctly identify positive instances. In the context of classification tasks, sensitivity provides insight into how effectively a model detects the class of interest, in this case, active molecules. A higher sensitivity score indicates that the model is proficient at identifying the positive class, which, in many domains, is crucial for downstream applications. Understanding sensitivity is particularly important in scenarios where the cost of missing a positive case (false negatives) outweighs the cost of misclassifying a negative one (false positives).

In this study, sensitivity was explored across various machine learning models applied to two different datasets: IRFMN—CERAPP and IRFMN-RBA. The results indicated notable differences in sensitivity, and these differences were deeply related to both the characteristics of the datasets and the intrinsic strengths and weaknesses of each model. The sensitivity values provided in Table 2 and Table 3 offer important insights into the relative strengths of different models in predicting molecular activity

3.3.3.1. Sensitivity in imbalanced datasets. The IRFMN—CERAPP dataset, being highly imbalanced, posed challenges for models in detecting the active molecules (positive class). Imbalanced datasets, where one class significantly outnumbers the other, can lead to models that are biased toward the majority class (inactive molecules in this case). This imbalance often results in models having low sensitivity, as they tend to predict the majority class more often and fail to identify many instances of the minority class.

The IRFMN—CERAPP 2D-QSAR model exhibited the highest sensitivity (0.85) compared to the 3D-QSAR models, such as kNN, RF, and

SVM. The 2D-QSAR approach focuses on extracting simple descriptors from the molecular structure that can be used to predict biological activity. While 2D descriptors may not capture the complex spatial interactions of molecules, they proved to be effective at identifying active molecules in this specific dataset. The model's ability to perform better in sensitivity highlights its robustness in recognizing the positive class despite the dataset's imbalance.

The RF model showed lower sensitivity in the IRFMN—CERAPP dataset (0.76), which can be attributed to its ensemble nature. RF builds decision trees using random subsets of features and data points, leading to diverse predictions. While RF models generally perform well, their sensitivity can suffer in imbalanced datasets because the algorithm might overly focus on the majority class during the construction of decision trees, especially if the random features selected for the trees do not adequately represent the minority class. Therefore, RF may miss many active molecules, which are often outnumbered by inactive ones in the IRFMN—CERAPP dataset.

The SVM algorithm constructs a decision boundary that maximizes the margin between the classes. However, when working with imbalanced datasets, SVM tends to perform poorly in terms of sensitivity because the decision boundary may lean heavily toward the majority class to minimize misclassifications. This results in a lower recall for the minority class, as the decision boundary is not flexible enough to account for the sparse distribution of positive instances. This is reflected in the lower sensitivity of SVM in the IRFMN—CERAPP dataset (0.82), where the model had difficulty classifying enough active molecules correctly.

Among the 3D-QSAR models, MLP achieved the highest sensitivity (0.89) on the IRFMN—CERAPP dataset. MLP, being a neural network, is well-suited for handling non-linear relationships and complex patterns in the data. Its multiple layers allow the model to learn intricate features and interactions, which improves its ability to detect positive instances. The higher sensitivity suggests that MLP was better able to classify the active molecules despite the dataset's imbalance, likely due to its ability to learn feature representations that distinguish the active from inactive molecules more effectively.

3.3.3.2. Sensitivity in balanced datasets. The IRFMN-RBA dataset, in contrast, is more balanced, with a more even distribution of active and inactive molecules. This balance significantly affects the performance of machine learning models, as the algorithms are not biased toward the majority class and can focus on detecting both active and inactive instances more effectively.

In this balanced dataset, the IRFMN-RBA 2D-QSAR model exhibited a sensitivity of 0.85, which was lower than that of the 3D-QSAR models but still respectable. The 2D-QSAR model again proved effective in identifying active molecules, demonstrating that while 2D-QSAR models may not capture the full molecular complexity, they are capable of achieving reasonable performance when the dataset is balanced.

For the IRFMN-RBA dataset, the RF model's sensitivity increased to 0.88, showcasing its improvement in a balanced dataset. The decision tree ensemble approach can now better handle both the positive and negative classes without being skewed by a majority class. RF's ability to perform better in a balanced dataset speaks to its inherent capacity for distinguishing between classes, making it more effective when the class distribution is more even.

In the case of the IRFMN-RBA dataset, SVM performed well with a sensitivity of 0.91. The more balanced dataset allowed SVM to construct a more accurate decision boundary between active and inactive molecules. SVM's ability to perform better with balanced datasets is due to its strategy of finding an optimal hyperplane that maximizes the margin between the two classes, a task that is more challenging when the dataset is imbalanced.

MLP also exhibited a high sensitivity of 0.91 for the IRFMN-RBA dataset, matching the performance of SVM. The neural network's

ability to capture complex patterns in the data is advantageous when the dataset is balanced, as it allows MLP to effectively identify both active and inactive molecules with high recall. The performance of MLP in this case is a strong indicator of its capability to adapt to various data distributions and still achieve superior results.

3.3.4. MCC of the models

MCC is a widely recognized binary classification metric that provides a more balanced measure of model performance than accuracy, particularly in imbalanced datasets. Unlike accuracy, which can be misleading when the dataset is skewed, MCC accounts for all elements of the confusion matrix—true positives, true negatives, false positives, and false negatives. This makes MCC especially valuable for evaluating models in situations where a high accuracy might result from predicting the majority class, without effectively classifying the minority class. For instance, in an imbalanced dataset with 90 % inactive and 10 % active molecules, a model predicting only inactive molecules would achieve high accuracy but fail to classify any active molecules. MCC, however, ensures that a high score is only achieved when both classes are predicted correctly [25], making it an ideal metric for datasets like IRFMN—CERAPP, which is highly imbalanced in favor of inactive molecules [26].

In the present study, MCC was calculated for several models developed on two datasets: IRFMN—CERAPP (highly imbalanced) and IRFMN-RBA (balanced). The results from the MCC scores, combined with accuracy, sensitivity, and specificity, provide a deeper understanding of the models' capabilities to correctly classify both positive and negative instances.

3.3.4.1. IRFMN—CERAPP dataset (Highly imbalanced). For the IRFMN—CERAPP dataset, MCC served as a more reliable indicator of model performance than accuracy, as the dataset is heavily imbalanced. Accuracy alone might suggest that a model performs well simply by classifying the majority class (inactive molecules), while it might fail to detect the minority class (active molecules) effectively.

From Table 2, it is evident that the 3D-QSAR models, except for KNN, consistently achieved higher MCC scores than the IRFMN—CERAPP 2D-QSAR model. The MLP model, which is a 3D-QSAR model, achieved the highest MCC of 0.87, reflecting its robust performance in both identifying active and inactive molecules correctly. This demonstrates the advantage of using sophisticated 3D-QSAR models over 2D-QSAR models, especially when dealing with imbalanced datasets, where a high MCC indicates a balanced prediction capability for both classes.

The KNN model, despite being a 3D-QSAR model, had a relatively lower MCC of 0.46, suggesting that it struggled with correctly classifying negative instances (inactive molecules) when compared to the other models. The KNN model's reliance on distance-based metrics might lead to difficulty in discerning subtle differences between molecules, especially when considering both classes in the imbalanced dataset.

The RF and SVM models also performed well with MCC values of 0.84, indicating that these models successfully identified both classes and showed robustness in performance, particularly in capturing the intricate patterns between molecular features.

Thus, the MCC values corroborate the accuracy results, revealing that 3D-QSAR models, especially MLP, provided a much better balance in correctly classifying both positive and negative instances than the 2D-QSAR model. This further supports the notion that MCC is more informative than accuracy in evaluating models for imbalanced datasets.

3.3.4.2. IRFMN-RBA dataset (Balanced). For the IRFMN-RBA dataset, which is balanced, MCC continues to offer a nuanced view of model performance. In Table 3, the MCC values for all models, particularly SVM and MLP, highlight the models' superior performance in correctly predicting both classes, with MCC values of 0.87 and 0.87, respectively. These values demonstrate that even when the dataset is balanced, MCC

is a valuable metric to ensure that the models are not biased toward one class.

Notably, the IRFMN-RBA 2D-QSAR model outperforms the KNN 3D-QSAR model in terms of MCC, with a score of 0.68 compared to 0.42 for KNN. This suggests that, even for balanced datasets, MLP and SVM are more adept at distinguishing between active and inactive molecules, with KNN showing less reliability, possibly due to its limitations in handling high-dimensional feature spaces effectively.

3.3.5. Model validation using external dataset

One of the key challenges in QSAR modeling is evaluating the predictive performance of models, and evaluation methodology has been the subject of many studies over the past several decades. The importance of model evaluation and comparison is reflected in the fourth principle of the Organization for Economic Cooperation and Development (OECD), which states that a QSAR model must have "appropriate measures of goodness of fit, robustness, and predictivity" [27]. While best practice guidelines often emphasize the need for external validation of compounds that have been rigorously excluded from the training set, implicit assumptions about errors in the training and validation data and how these assumptions might affect performance evaluation tend to be overlooked [28].

In that regard, accuracy and sensitivity are considered measures for comparing and validating the robustness of 3D-QSAR models over 2D-QSAR models used in VEGA. The external dataset for IRFMN-RBA consisted of only active molecules; hence, the specificity and MCC were not calculated. As shown in Fig. 9, the accuracies of 3D-QSAR models such as SVM, RF, and MLP were better than that of IRFMN—CERAPP on external dataset; however, IRFMN—CERAPP was highly sensitive. Similarly, as shown in Fig. 10, the accuracy and sensitivity of the MLP 3D-QSAR model were better than those of IRFMN-RBA on external dataset, while IRFMN-RBA was superior to the other 3D-QSAR models.

4. Conclusion

To conclude, the results of 3D-QSAR modeling, combined with various cheminformatics techniques, highlight the importance of proper data partitioning to avoid overfitting or underfitting. Machine learning-based algorithms proved essential for developing 3D-QSAR models, offering significant advantages over 2D-QSAR models, particularly in predicting the estrogen receptor-mediated effects of hER α . The hER α gene, a key target in endocrine disruption, was effectively modeled, with the MLP 3D-QSAR model emerging as a robust tool for predicting the binding affinity and activity of chemically diverse estrogen receptor molecules. These findings underscore the value of dataset optimization and advanced modeling techniques for improving predictive accuracy in molecular activity and binding studies.

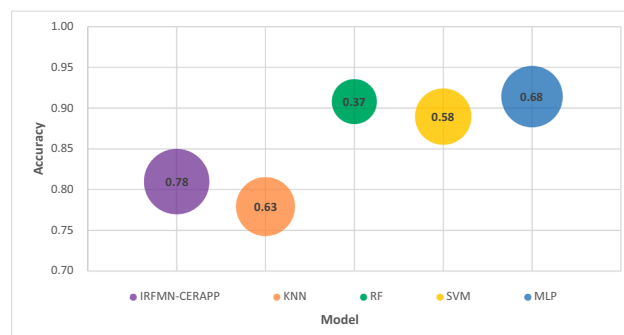


Fig. 9. The plot illustrating the predictive ability and the robustness of different QSAR models developed for IRFMN—CERAPP dataset to predict the potential of small molecule in the external dataset for hER α binding activity. The radius of the circle corresponds to sensitivity and each model is represented in different colors.

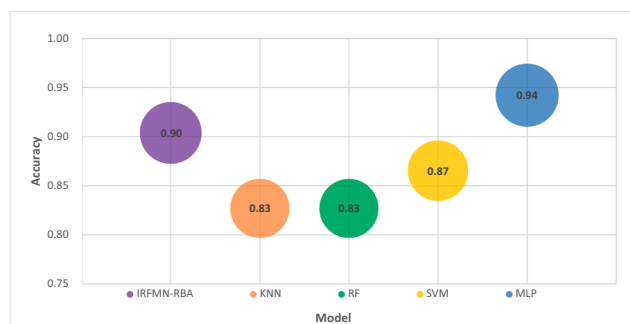


Fig. 10. The plot illustrating the predictive ability and the robustness of different QSAR models developed for IREFMN-RBA dataset to predict the potential of small molecule in the external dataset for hER α binding activity. The radius of the circle corresponds to sensitivity and each model is represented in different colors.

CRediT authorship contribution statement

BR. Bharath: Writing – original draft, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Sreerupa Mitra:** Software, Resources, Methodology, Conceptualization. **Nadeem Khan:** Writing – review & editing, Supervision, Project administration.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Bharath BR reports financial support, administrative support, statistical analysis, and writing assistance were provided by Jai Research Foundation. Bharath BR reports a relationship with Jai Research Foundation that includes: employment. Bharath BR reports a relationship with Jai Research Foundation that includes: employment. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at [doi:10.1016/j.ins.2025.100014](https://doi.org/10.1016/j.ins.2025.100014).

Data availability

Data will be made available on request.

References

- [1] Shanle EK, Xu W. Endocrine disrupting chemicals targeting estrogen receptor signaling: identification and mechanisms of action. *Chem Res Toxicol* 2011;24: 6–19. <https://doi.org/10.1021/tx100231n>.
- [2] Hall JM, Couse JF, Korach KS. The multifaceted mechanisms of estradiol and estrogen receptor signaling. *J Biol Chem* 2001;276:36869–72. <https://doi.org/10.1074/jbc.r100029200>.
- [3] Blair RM, Fang H, Branham WS, Hass BS, Dial SL, Moland CL, et al. The estrogen receptor relative binding affinities of 188 natural and xenochemicals: structural diversity of ligands. *Toxicol Sci* 2000;54:138–53. <https://doi.org/10.1093/toxsci/54.1.138>.
- [4] Schug TT, Janesick A, Blumberg B, Heindel JJ. Endocrine disrupting chemicals and disease susceptibility. *J Steroid Biochem Mol Biol* 2011;127:204–15. <https://doi.org/10.1016/j.jsbmb.2011.08.007>.
- [5] Roncaglioni A, Piclin N, Pintore M, Benfenati E. Binary classification models for endocrine disrupter effects mediated through the estrogen receptor. *SAR QSAR Environ Res* 2008;19:78. <https://doi.org/10.1080/10629360802550606>.
- [6] Danieli A, Colombo E, Raitano G, Lombardo A, Roncaglioni A, Manganaro A, Sommovigo A, Carnesecchi E, Dorne JCM, Benfenati E. The VEGA tool to check the applicability domain gives greater confidence in the prediction of in silico models. *Int J Mol Sci* 2023;24(12):9894. <https://doi.org/10.3390/ijms24129894>.
- [7] Zhu H, Zhang J, Kim MT, Boison A, Sedykh A, Moran K. Big data in chemical toxicity research: the use of high-throughput screening assays to identify potential toxicants. *Chem Res Toxicol* 2014;27:1643–51. <https://doi.org/10.1021/tx500145h>.
- [8] Mansouri K, Abdelaziz A, Rybacka A, et al. CERAPP: collaborative estrogen receptor activity prediction project. *Environ Health Perspect* 2016;124(7): 1023–33. <https://doi.org/10.1289/ehp.1510267>.
- [9] Sato A, Miyao T, Jasial S, et al. Comparing predictive ability of QSAR/QSPR models using 2D and 3D molecular representations. *J Comput Aided Mol Des* 2021;35: 179–93. <https://doi.org/10.1007/s10822-020-00361-7>.
- [10] Zhang X, Mao J, Li W, Koike K, Wang J. Improved 3D-QSAR prediction by multiple-conformational alignment: a case study on PTP1B inhibitors. *Comput Biol Chem* 2019;83:107134. <https://doi.org/10.1016/j.compbiolchem.2019.107134>.
- [11] De Simone A, Russo D, Ruda GF, Micoli A, Ferraro M, Di Martino RMC, Cavalli A. Design, synthesis, structure-activity relationship studies, and three-dimensional quantitative structure-activity relationship (3D-QSAR) modeling of a series of OBiphenyl carbamates as dual modulators of dopamine D3 receptor and fatty acid amide hydrolase. *J Med Chem* 2017;60:2287–304.
- [12] Uddin R, Naz A, Akhtar N, ul Haq Z. Development of robust QSAR model using rapid overlay of crystal structures (ROCS) based alignment: a test case of Tubulin inhibitors. *Med Chem Res* 2012;22:3229–41. <https://doi.org/10.1007/s00044-012-0327-0>.
- [13] Nezirina K, Nenad T, Sanja M, Elenora P, Manuela S, Lorenzo A, Milan M, Rino R. Human estrogen receptor alpha antagonists, part 3: 3-D pharmacophore and 3-D QSAR guided brefeldin A hit-to-lead optimization toward new breast cancer suppressants. *Molecules* 2022;27(9). <https://doi.org/10.3390/molecules27092823>. 2823–2823.
- [14] Wenliang J, Qinghua C, Bo Z, Fangfang W. In silico prediction of estrogen receptor subtype binding affinity and selectivity using 3D-QSAR and molecular docking. *Med Chem Res* 2019;28(11):1974–94. <https://doi.org/10.1007/s00044-019-02428-z>.
- [15] Pedretti A, Mazzolari A, Gervasoni S, Fumagalli L, Vistoli G. The VEGA suite of programs: a versatile platform for cheminformatics and drug design projects. *Bioinformatics* 2020;37:1174–5. <https://doi.org/10.1093/bioinformatics/btaa774>.
- [16] Joseph VR. Optimal ratio for data splitting. *Statistical analysis and data mining: the ASA. Data Sci J* 2022. <https://doi.org/10.1002/sam.11583>.
- [17] Michael MG. Chapter 7 - protein interactions, protein bioinformatics. *Academic Press*; 2010.
- [18] Yang ZD, Yu J, Zhang Q. Effects of raloxifene on cognition, mental health, sleep and sexual function in menopausal women: a systematic review of randomized controlled trials. *Maturitas* 2013;75(4):341–8. <https://doi.org/10.1016/j.maturitas.2013.05.010>.
- [19] Choi Y, Deane CM. FREAD revisited: accurate loop structure prediction using a database search algorithm. *Proteins* 2010;78:1431–40. <https://doi.org/10.1002/prot.22658>.
- [20] Bharath BR, Abhishek KT, Vaibav B, Nitin P, Ashita D, Sreerupa M, Abhay D. Development and experimental validation of 3D QSAR models for the screening of thyroid peroxidase inhibitors using integrated methods of computational chemistry. *Heliyon* 2024;10(8). <https://doi.org/10.1016/j.heliyon.2024.e29756>. E29756.
- [21] Damle L, Damle H, Bharath BR. Plant formulation ATRICOV 452 in improving the level of COVID-19 specific inflammatory markers in patients. *Contemp Clin Trials Commun* 2022;28:100961. <https://doi.org/10.1016/j.conctc.2022>.
- [22] Aziz M, Ejaz SA, Zargar S, Akhtar N, Aborode AT, Wani T, Batiha GE, Siddique F, Alqarni M, Akintola AA. Deep learning and structure-based virtual screening for drug discovery against NEK7: a novel target for the treatment of cancer. *Molecules* 2022;27(13):4098. <https://doi.org/10.3390/molecules27134098>.
- [23] Valsecchi C, Grisoni F, Consonni V, Ballabio D. Consensus versus individual QSARs in classification: comparison on a large-scale case study. *J Chem Inf Model* 2020;60(3):1215–23. <https://doi.org/10.1021/acs.jcim.9b01057>.
- [24] Wilm A, Stork C, Bauer C, Schepky A, Kühnl J, Kirchmair J. Skin doctor: machine learning models for Skin sensitization prediction that provide estimates and indicators of prediction reliability. *Int J Mol Sci* 2019;20(19):4833. <https://doi.org/10.3390/ijms20194833>.
- [25] Chicco D. Ten quick tips for machine learning in computational biology. *BioData Min* 2017;10(35):1–17. <https://doi.org/10.1186/s13040-017-0155-3>.
- [26] He H, Garcia EA. Learning from imbalanced data. *IEEE Trans Knowl Data Eng* 2009;21(9):1263–84. <https://doi.org/10.1109/TKDE.2008.239>.
- [27] OECD principles for the validation, for regulatory purposes, of (quantitative) structure-activity relationship models. <https://www.oecd.org/chemicalsafety/risk-assessment/37849783.pdf> [Accessed 18 March 2024].
- [28] Tropsha A. Best practices for QSAR model development, validation, and exploitation. *Mol Inf* 2010;29. 476–88.s.