Title

Predicting the binding of small molecules to nuclear receptors using machine learning

Journal

Briefings in Bioinformatics, 23(3)

ISSN

1467-5463

Authors

Ramaprasad, Azhagiya Singam Ettayapuram

Smith, Martyn T

McCoy, David

et al.

Publication Date

2022-05-13

DOI

10.1093/bib/bbac114

Peer reviewed

OXFORD

Briefings in Bioinformatics, 2022, **23(3)**, 1–13
https://doi.org/10.1093/bib/bbac114
Problem Solving Protocol

Predicting the binding of small molecules to nuclear receptors using machine learning

Azhagiya Singam Ettayapuram Ramaprasad, Martyn T. Smith, David McCoy, Alan E. Hubbard, Michele A. La Merrill and Kathleen A. Durkin

Corresponding author. Kathleen A. Durkin, Molecular Graphics and Computation Facility, College of Chemistry, University of California, Berkeley, CA 94720, USA. Tel.: +1(510) 642-6719; E-mail: durkin@berkeley.edu

Abstract

Nuclear receptors (NRs) are important biological targets of endocrine-disrupting chemicals (EDCs). Identifying chemicals that can act as EDCs and modulate the function of NRs is difficult because of the time and cost of *in vitro* and *in vivo* screening to determine the potential hazards of the 100 000s of chemicals that humans are exposed to. Hence, there is a need for computational approaches to prioritize chemicals for biological testing. Machine learning (ML) techniques are alternative methods that can quickly screen millions of chemicals and identify those that may be an EDC. Computational models of chemical binding to multiple NRs have begun to emerge. Recently, a Nuclear Receptor Activity (NuRA) dataset, describing experimentally derived small-molecule activity against various NRs has been created. We have used the NuRA dataset to develop an ensemble of ML-based models to predict the agonism, antagonism, binding and effector binding of small molecules to nine different human NRs. We defined the applicability domain of the ML models as a measure of Tanimoto similarity to the molecules in the training set, which enhanced the performance of the developed classifiers. We further developed a user-friendly web server named 'NR-ToxPred' to predict the binding of chemicals to the nine NRs using the best-performing models for each receptor. This web server is freely accessible at http://nr-toxpred. cchem.berkeley.edu. Users can upload individual chemicals using Simplified Molecular-Input Line-Entry System, CAS numbers or sketch the molecule in the provided space to predict the compound's activity against the different NRs and predict the binding mode for each.

Keywords: Nuclear Receptor, Toxicity, Super Learner, Machine learning

Introduction

Nuclear receptors (NRs) are a large family of transcription factors that can be activated by lipophilic ligands and bind directly to DNA to regulate the expression of target genes [1–3]. In response to metabolic and endocrine ligands such as gonadal and adrenal steroids, NRs are fundamental regulators of cellular responses that impact many aspects of development, reproduction and metabolism [4, 5]. Therefore, they are important targets for pharmaceuticals, and as such, ∼13% of US Food and Drug Administration-approved drugs target NRs [6]. They are also important targets for the toxic effects of many xenobiotic chemicals, which can act as agonists or antagonists to the receptor in question. For example, one of the key characteristics of carcinogens is their ability to modulate NR-mediated effects [7]. The two key characteristics of endocrine-disrupting chemicals (EDCs) are the activation or antagonism of nuclear hormone receptors [8].

As many as 350 000 chemicals and chemical mixtures are registered for use in everyday commerce [9], many of which have not been tested for toxic effects, including carcinogenic and endocrine-disrupting activity. The considerable number of existing chemicals coupled with the diversity of potential NR targets means that experimental testing of NR binding has sparsely covered this multidimensional space. Although experimental protocols exist to test if chemicals interact with various NRs, the applications of these *in vivo* and *in vitro* techniques are severely limited by the cost and time per chemical analy-

Azhagiya Singam Ettayapuram Ramaprasad is a Research Specialist in the Molecular Graphics and Computation Facility at the University of California, Berkeley. His work focuses on the application of computational biology/chemistry towards understanding small molecule interactions with protein structures and predicting toxicity. He holds a PhD in Bioinformatics/Computational Biology from the University of Madras.
Martyn T. Smith is a Distinguished Professor of Toxicology and the Kaiser Chair of Cancer Epidemiology in the School of Public Health at the University of California, Berkeley where he is also the Director of the Superfund Research Program.
David McCoy is a PhD student in the School of Public Health at the University of California, Berkeley where he works on machine learning algorithms with applications in epidemiology and environmental toxicology. He has extensive algorithm development experience and has returned to academia after several years of work in the commercial technology sector.
Alan E. Hubbard is a Professor of Biostatistics at the University of California, Berkeley where he works on estimation of complex causal parameters and prediction algorithms using machine learning, with an emphasis on applications in epidemiology, environmental exposure and biomedicine.
Michele A La Merrill is an Associate Professor in Environmental Toxicology at the University of California, Davis. She holds a PhD in toxicology from the University of North Carolina at Chapel Hill and a MPH in epidemiology from the Mount Sinai School of Medicine.
Kathleen A. Durkin has been the Director of the Molecular Graphics and Computation Facility at the University of California, Berkeley for more than 25 years. She holds a PhD in Chemistry from Emory University and a BS in Oceanography from the Florida Institute of Technology.
Received: December 13, 2021. **Revised:** March 7, 2022. **Accepted:** March 9, 2022

sis. Further, the addition of novel chemicals to commerce is ongoing, and predictive models could better ensure the development of safer chemicals for a circular economy. Given the importance of NRs as molecular targets, it would be of great benefit if the binding of any given chemical against a particular NR could be predicted.

Most prior research on the development of various computational models for predicting and prioritizing chemicals for their potential to interact with NRs has focused on one NR at a time [10–13]. For example, Maran et al. have utilized random forest-based models to classify chemicals for their ability to bind to the androgen receptor (AR) [14]. Singam et al. utilized both molecular docking and machine learning (ML) to screen over 5000 per- and poly-fluorinated alkyl substances and chemicals associated with hydraulic fracturing for their ability to bind human AR [15, 16]. Researchers have utilized molecular docking simulation alone for predictive classification of AR, estrogen receptor (ER) and pregnane X receptor (PXR) binding [17–19]. Other investigators have focused on developing a combination of computational methods, such as docking and quantitative structure-activity relationship (QSAR), and generated consensus models for characterizing chemical binding properties against various additional NRs such as ER, AR, peroxisome proliferator activated receptor gamma (PPARG) and progesterone receptor (PR) [10, 11, 20–22].

Computational models of chemical binding to multiple NRs have begun to emerge. Recently, Ballabio and coworkers have created a comprehensive dataset of chemicals called the Nuclear Receptor Activity (NuRA) dataset [23], describing experimentally derived small-molecule activity against various NRs by curating different databases, including the Toxicology in the 21st Century (Tox21) [24] program database, ChEMBL [25], BindingDB [26] and NR-DBIND [27]. Valsecchi *et al.* [28] have also developed models using the NuRA dataset and compared the classification performance of multitask deep neural networks with single-task benchmark classification approaches in a QSAR perspective. In this study, instead of training a single ML algorithm to model molecular features for different NRs, we utilized an ensemble approach (Super learner) to fit a prediction function for each of our outcomes. No previous study has developed a similar prediction tool by combining such rich high-dimensional features with cross-validated optimization among many competing algorithms for such a wide range of NRs.

EDCs can interfere with biological processes such as hormone synthesis, metabolism, homeostasis, reproduction, development, etc. Identifying chemicals that can act as EDCs is difficult owing to the time and cost of *in vitro* and *in vivo* screening of the 100 000s of chemicals that humans are exposed to. Computational prediction of chemicals with activity against these NRs will help prioritize chemicals for further *in vitro* and *in vivo* evaluation, subsequent regulation and safer chemical replacement using green chemistry. As well-curated NuRA datasets for nine receptors are available, herein a systematic attempt has been made to develop ML-based models using chemical fingerprints to predict the chemical binding of nine different NRs: AR, ERA, ERB, farnesoid X receptor (FXR), glucocorticoid receptor (GR), PR, peroxisome proliferator-activated receptor delta (PPARD), PPARG and retinoid X receptor (RXR), using the NuRA chemical dataset for model training. We also developed a web-based application using our computational models to enable *de novo* prediction of the binding of small molecules to these nine different NRs. Furthermore, we deployed a publicly available, user-friendly web server that allows the experimental and regulatory community to screen for NR-binding chemicals, on demand, as new chemicals are produced or become prominent in their application.

Materials and Methods

Dataset

This study used the well-curated NuRA chemical dataset to train ML models for nine NRs [23]. The dataset and the Konstanz Information Miner (KNIME) workflow [29] of data curation were downloaded from https://doi.org/10.5281/zenodo.3991561. We carefully verified each step of the curation process. The dataset contains 15 247 combined entries for nine different receptors, annotated as three binding class types: (i) agonist, (ii) antagonist and (iii) binders. Each type is further classified as activity type (i) active, (ii) weakly active, (iii) inactive, (iv) inconclusive and (v) data missing. Table 1 shows the compositions of different classes for each receptor. Missing data and inconclusive results were removed from the dataset. Then, because the number of chemicals in the weakly active category is low, we combined the active and weakly active entries into a single category in each binding class type, resulting in a binary (active versus inactive) designation for each of the agonists, antagonists and binders. Our study, therefore, developed ML models to predict each of these binding class types using a binary classification (binding class models).

As we are also interested in identifying active binding versus inactive chemicals (effector models) regardless of agonist, antagonist or undefined binding class, we additionally developed ML models by first merging the three binding classes and removing the inconclusive and missing data for each receptor to increase the sample size of effector types (actives and inactive). Table 2 shows the active and inactive chemical compositions for each receptor after merging the three binding types.

Training dataset

For each of the nine NRs, the NR-specific curated chemical datasets were randomly divided into training (80%) and validation sets (20%) using the 'train_test_split' function in the scikit-learn package (Tables 1 and 2). The validation set was used to give an estimate of the performance of each of the developed models. This 20% validation set of chemicals was not used in the training

Table 1. Number of chemicals by class for all receptors in the training and validation set

Receptor	Class	Total inactive	Total active	Total weakly active	Training set actives/inactives	Validation set actives/inactives
	Agonist	5670	349	27	290/4546	86/1124
PR	Antagonist	4400	741	548	1027/3524	262/876
	Binder	5040	1251	53	1057/4018	247/1022
	Agonist	4549	130	133	–	–
RXR	Antagonist	3	115	1	–	–
	Binder	4569	861	145	–	–
	Agonist	5384	737	41	613/4316	165/1068
GR	Antagonist	4577	657	190	666/3673	181/904
	Binder	5228	1815	84	1537/4164	362/1064
	Agonist	5578	513	121	517/4452	117/1126
AR	Antagonist	4942	776	391	926/3961	241/981
	Binder	5130	1419	104	1243/4079	280/1051
	Agonist	5060	476	461	751/4046	186/1014
ERA	Antagonist	5160	362	322	544/4131	140/1029
	Binder	4861	1287	177	1184/3876	280/985
	Agonist	5744	286	48	270/4592	64/1152
ERB	Antagonist	5133	224	229	359/4109	94/1024
	Binder	5554	1159	66	998/4425	227/1129
	Agonist	5349	372	85	346/4298	111/1051
FXR	Antagonist	4829	124	143	219/3857	48/972
	Binder	5272	550	108	530/4214	128/5272
	Agonist	5663	616	73	–	–
PPARD	Antagonist	5561	28	24	–	–
	Binder	5742	730	52	–	–
	Agonist	5223	1352	158	1200/4186	310/1037
PPARG	Antagonist	5249	88	153	203/4189	38/1060
	Binder	5458	1699	205	1529/4360	375/1098

Table 2. Number of active and inactive chemicals for all receptors

	Total		Training set		Validation set	
Receptor	Actives	Inactives	Total	Actives/inactives	Total	Actives/inactives
RXR	1008	4569	4461	807/3654	1116	201/915
PR	2078	5063	5712	1646/4066	1429	432/997
GR	2143	5232	5900	1720/4180	1475	423/1052
AR	2217	5179	5916	1782/4134	1480	435/1045
ERA	2327	4956	5826	1863/3963	1457	464/993
ERB	1552	5563	5692	1228/4464	1423	324/1099
FXR	837	5276	4890	662/4228	1223	175/1048
PPARD	848	5745	5274	678/4596	1319	170/1149
PPARG	2118	5469	6069	1693/4376	1518	425/1093

process while developing and optimizing any of our ML models.

Molecular features

In this investigation, we utilized molecular fingerprints for descriptor features. We employed two widely used fingerprinting methods: (i) Morgan fingerprints, also called extended-connectivity fingerprints (ECFP4), which are circular substructure fingerprints where we chose a radius of 3 and a length of hashed binary vectors of 1024-bits; and (ii) Molecular Access System (MACCS) fingerprints, which have 166 public keys implemented as SMILES arbitrary target specification (SMARTS). The Python-based RDKit [30] library was used to generate the molecular fingerprints from the Simplified Molecular-Input Line-Entry System (SMILES) data.

ML model development

As noted previously [31], there is no single optimal ML algorithm for all potential data problems. However, one can define an approach that is guaranteed to generate the best from a set of explicit, competing algorithms. In our case, we used nine different ML techniques, including (i) AdaBoost [32], which is a boosting algorithm that combines multiple 'weak classifiers' into a single 'strong classifier'; (ii) logistic regression [33], which predicts the value of a categorical variable based on its relationship with predictor variables; (iii) random forest [34],

which merges a collection of independent decision trees to decrease both bias and variance; (iv) support vector machine (SVM) [35], which is a classifier that finds an optimal hyperplane to maximize the margin between two classes; (v) k-nearest neighbors (k-NN) algorithm [36], which assumes that similar data points exist near each other and makes predictions by calculating the difference between the new data point and all other data points in the training set; (vi) Bagging classifier [37], which is an ensemble-based model that fits base classifiers on random subsets of the original dataset and then aggregates their predictions to generate a final prediction; (vii) Gaussian naive Bayes [38], which is a variant of the naive Bayes algorithm based on Bayes theorem; (viii) decision tree classifier algorithm [39], which uses a tree where each node represents a feature, each branch represents the decision and each leaf represents an outcome and (ix) Super learner [31], which combines the predictive probabilities of NR binding across many ML algorithms and finds the optimal combination of the collection of algorithms by minimizing the cross-validated risk. This approach is an improvement over methods using only one ML algorithm because no one algorithm is universally optimal. Super learner has been shown in theory to be at least as good as the best performing algorithm in the ensemble and often performs considerably better than the component ML models. For each of these methods, we used a grid-search cross-validation (CV) method (GridSearchCV) as implemented in scikit-learn [40] to tune the hyperparameters.

Repeated k-fold CV

We assessed the performance of the classification models using stratified k-fold CV. The stratified-folds function was utilized to split the data while keeping the correct ratio of different classes. We evaluated the classification performance for each receptor by repeated stratified k-fold CV with ten splits and 100 repeats, thus in total 1000-fold.

Evaluating ML models

The performance of the developed models was evaluated using sensitivity, specificity, balanced accuracy and Matthew's correlation coefficient (MCC) [41]. The dataset used in this study is imbalanced because the number of active binders is less than the inactive chemicals. For imbalanced data classification, the receiver operating characteristic (ROC) area under the curve (AUC) can be misleading because a small number of correct or incorrect predictions can result in a large change in the ROC curve [42]. Hence, the standard alternative precision-recall curve (PR AUC) is used to assess the performance of the ML models. The balanced accuracy metric deals with the imbalance in datasets, for example, where actives are not equal in number to inactives. It weights each sample's raw accuracy according to the inverse of its actual prevalence [43]**.** The following equations were used to calculate sensitivity, specificity and MCC [41]:

$$\text{Sensitivity} = \frac{\text{TP}}{(\text{TP} + \text{FN})} \tag{1}$$

$$\text{Specificity} = \frac{\text{TN}}{(\text{TN} + \text{FP})} \tag{2}$$

$$\text{MCC} = \frac{(\text{TP})\,(\text{TN}) - (\text{FP})\,(\text{FN})}{\sqrt{(\text{TP} + \text{FN})\,(\text{TN} + \text{FP})\,(\text{TP} + \text{FP})\,(\text{TN} + \text{FN})}}, \tag{3}$$

where TP, FP, FN and TN are true positive, false positive, false negative and true negative, respectively.

Applicability domain

The applicability domain is defined as described by Chen *et al.* [44] and was measured by the similarity to the molecules in the training set. Tanimoto similarity was calculated using ECFP4 and MACCS fingerprints for the respective feature spaces. The test molecule is considered to be within the applicability domain if the number of chemicals [N_{min} (default = 1)] with similarity is greater than the cutoff [S_{cutoff} (default = 0.25)] in the training dataset. The applicability domain was defined as a combination of S_{cutoff} and N_{min}.

Web server implementation

The user interface of the NR-ToxPred web server was developed using the Python Django framework. J(s)mol, an open-source JavaScript library, was used in the front-end web application. The RDKit package [30] (http://www.rdkit.org/) is used for processing the input SMILES string. NR-ToxPred is built on an Apache HTTP server at the back end. Redis is employed to queue and assess the application programming interface (API) requests. Users can input a SMILES string or draw 2D structures using a J(s)mol interface [45]. Users also have an option to submit multiple SMILES entries in comma-separated values (CSV) formatted files. They can analyze up to ~5000 chemicals at a time. A sample CSV file is available on the webserver for formatting guidance. Users can select the applicability domain N_{min} and S_{cutoff} settings to check for the model's reliability to the given molecule. The NR-ToxPred prediction results are tabulated for each receptor along with their scores. As it is also essential to understand the atomic-level details of the ligand binding to the NR, we also implemented a molecular docking protocol using Autodock vina [46]. This docking is run on a given input chemical to the target NR if the ML model predicts the input structure as an active ligand. Suppose the docking score is unfavorable at this step. In that case, the results suggest that the ligand may not fit in the ligand-binding pocket. This process can serve as a check on any false positives in the ML output. Users can visualize or download the ligand and receptor docking coordinates on the results page or receive an email with the results.

Results

Hyperparameter tuning

We evaluated the performance of nine different ML algorithms, each with two different fingerprints as features. Hyperparameters were tuned using grid search 10-fold stratified CV. The optimized hyperparameters for each algorithm with ECFP4 fingerprints and MACCS fingerprints for all the receptors for effector models are given in supporting information Table S1 (see Supplementary Data available online at *Briefings in Bioinformatics*) and binding class models in supporting information Table S2 (see Supplementary Data available online at *Briefings in Bioinformatics*). After tuning the hyperparameters for each algorithm, we trained each ML model and evaluated it by CV methods and with the validation set.

Models for AR

Binding class models for AR

Agonist, antagonist and binder datasets were used to build three ML models for AR. Prediction accuracy for different types and algorithms on repeated stratified k-fold CV with ECFP4 and MACCS fingerprints is given in Tables 3 and 4, respectively. The different algorithms on the agonist and binder dataset have achieved a stratified k-fold CV prediction accuracy of >90%. The best accuracy was obtained with both Super learner and SVM-based models for the agonist dataset: 87% on the validation set with ECFP4 fingerprints (Table 5). With the MACCS fingerprints, the best accuracy for the agonist dataset was obtained for Super learner (Table 6). Both SVM and Super learner had similar performance measures with 97 and 96% accuracy on the validation set for ECFP4 and MACCS fingerprints for the binder dataset. For the agonist dataset, the PR AUC values of the validation dataset for Super learner and SVM were 0.81 and 0.80 (Table 5), respectively, for ECFP4 fingerprints and 0.81 and 0.79 for MACCS fingerprints (Table 6). The validation dataset's PR AUC value is 0.98 and 0.97 for ECFP4 and MACCS fingerprints for the binder dataset.

For the antagonist model, AdaBoost classifier, Bagging classifier, decision tree classifier, k-NN, random forest, Super learner and SVM models have achieved a prediction accuracy of >85% with both ECFP4 and MACCS fingerprints as a feature. On the validation set with ECFP4 fingerprints, Super learner and SVM-based models achieved 83 and 84% accuracy, respectively (Table 5). Similar balanced accuracy was obtained for Super learner and SVM models with MACCS fingerprints (Table 6). The PR AUC values on the validation dataset for Super learner and SVM are 0.81 and 0.80 (Table 5), respectively, for ECFP4 fingerprints and 0.81 and 0.79 for MACCS fingerprints (Table 6). The developed model's performance is comparable to other developed models [11, 28].

Effector models for AR

For AR, four algorithms (k-NN, random forest, SVM and Super learner) with ECFP4 fingerprints all exhibited high predictive power. The balanced accuracy values are 85, 86, 87 and 86%, respectively, with MCC scores of 0.77, 0.73, 0.78 and 0.89, respectively, on the validation dataset (Table 7). The accuracy scores on the repeated stratified k-fold CV for these three models are 0.90 ± 0.01, 0.88 ± 0.01, 0.89 ± 0.01 and 0.90 ± 0.01 (Table 8). The effector AR model has achieved a prediction accuracy of 90% on the repeated stratified k-fold CV for SVM and k-NN and 89% for Super learner. Although k-NN and SVM achieved higher accuracy with MACCS fingerprints, SVM with ECFP4 fingerprints performed best with a higher MCC value, which produced a more informative and truthful score in evaluating binary classifications [47].

Models for Estrogen Receptor Alpha (ERA) and Estrogen Receptor Beta (ERB)

Binding class models for ERA and ERB

ML models of agonist, antagonist and binder of both ERA and ERB were evaluated using the validation dataset and repeated stratified k-fold CV. The performance measures for different algorithms with the validation set and repeated stratified k-fold CV are given in Tables 3 and 4 for ECFP4 and MACCS fingerprints as input features, respectively. The Bagging classifier has an average accuracy of 89, 91 and 94% for agonist, antagonist and binder datasets with ECFP4 fingerprints and 88, 91 and 93% with MACCS fingerprints, respectively, for ERA. The performance measure for ERA and ERB datasets using the binding class classifier on the validation set are given in supporting information Tables S3 and S4 (see Supplementary Data available online at *Briefings in Bioinformatics*), respectively, for ECFP4 fingerprints as input feature and Tables S5 and S6 (see Supplementary Data available online at *Briefings in Bioinformatics*), respectively, for MACCS fingerprints. Even though the Bagging classifier has better accuracy on repeated stratified k-fold CV, SVM and Super learner appear to give more consistent prediction accuracies on both repeated stratified k-fold CV and the validation dataset (Tables S3 and S5**,** see Supplementary Data available online at *Briefings in Bioinformatics*). Similarly, for ERB, more consistent performance measures were obtained with SVM and Super learner (see Tables S4 and S6**,** see Supplementary Data available online at *Briefings in Bioinformatics*).

Effector models for ERA and ERB

The SVM model performed best (balanced accuracy, 80%; MCC score of 0.66), followed by Random forest (accuracy, 79%; MCC score 0.61) for ECFP4 fingerprints as descriptors on the validation dataset of ERA (Table S7**,** see Supplementary Data available online at *Briefings in Bioinformatics*). For MACCS fingerprints, SVM had comparable accuracy but a lower MCC (Table S7**,** see Supplementary Data available online at *Briefings in Bioinformatics*). The lower MCC was likely due to the promiscuous nature of ERA, which binds to diverse chemicals, which in turn made it somewhat harder for ML algorithms to discriminate between NR-binding and nonbinding chemicals. For ERB, the accuracy score on the repeated stratified k-fold CV is 85 and 86% for Super learner and SVM for ECFP4 fingerprints (Table S8**,** see Supplementary Data available

Table 3. Average accuracy of different algorithms for three class approach for six receptors using ECFP4 fingerprints as input features on the repeated k-fold CV

Receptor	Algorithm	Agonist Accuracy	Antagonist Accuracy	Binder Accuracy
AR	AdaBoost classifier	0.95 ± 0.01	0.88 ± 0.01	0.96 ± 0.01
	Bagging classifier	0.96 ± 0.01	0.89 ± 0.01	0.94 ± 0.01
	Decision tree classifier	0.95 ± 0.01	0.88 ± 0.01	0.95 ± 0.01
	Gaussian naive Bayes	0.92 ± 0.01	0.83 ± 0.02	0.91 ± 0.01
	k-NN	0.93 ± 0.01	0.88 ± 0.01	0.98 ± 0.01
	Logistic regression	0.92 ± 0.01	0.84 ± 0.01	0.96 ± 0.01
	Random forest	0.95 ± 0.01	0.86 ± 0.01	0.96 ± 0.01
	Super learner	0.95 ± 0.01	0.87 ± 0.02	0.98 ± 0.01
	SVM	0.96 ± 0.01	0.88 ± 0.01	0.98 ± 0.01
ERA				
	AdaBoost classifier	0.86 ± 0.01	0.90 ± 0.01	0.93 ± 0.01
	Bagging classifier	0.89 ± 0.01	0.91 ± 0.01	0.94 ± 0.01
	Decision tree classifier	0.87 ± 0.01	0.91 ± 0.01	0.93 ± 0.01
	Gaussian naive Bayes	0.81 ± 0.02	0.81 ± 0.02	0.88 ± 0.01
	k-NN	0.80 ± 0.02	0.90 ± 0.01	0.94 ± 0.01
	Logistic regression	0.78 ± 0.02	0.86 ± 0.01	0.93 ± 0.01
	Random forest	0.86 ± 0.01	0.85 ± 0.02	0.94 ± 0.01
	Super learner	0.84 ± 0.02	0.86 ± 0.02	0.94 ± 0.01
	SVM	0.84 ± 0.01	0.88 ± 0.01	0.94 ± 0.01
ERB				
	AdaBoost classifier	0.97 ± 0.01	0.93 ± 0.01	0.97 ± 0.01
	Bagging classifier	0.98 ± 0.01	0.94 ± 0.01	0.97 ± 0.01
	Decision tree classifier	0.97 ± 0.01	0.92 ± 0.01	0.96 ± 0.01
	Gaussian naive Bayes	0.95 ± 0.01	0.84 ± 0.02	0.93 ± 0.01
	k-NN	0.95 ± 0.01	0.92 ± 0.01	0.97 ± 0.01
	Logistic regression	0.97 ± 0.01	0.89 ± 0.01	0.97 ± 0.01
	Random forest	0.96 ± 0.01	0.89 ± 0.01	0.97 ± 0.01
	Super learner	0.96 ± 0.01	0.86 ± 0.02	0.98 ± 0.01
	SVM	0.96 ± 0.01	0.87 ± 0.01	0.98 ± 0.01
FXR				
	AdaBoost classifier	0.97 ± 0.01	0.95 ± 0.01	0.96 ± 0.01
	Bagging classifier	0.98 ± 0.01	0.96 ± 0.00	0.96 ± 0.01
	Decision tree classifier	0.97 ± 0.01	0.94 ± 0.01	0.96 ± 0.01
	Gaussian naive Bayes	0.96 ± 0.01	0.88 ± 0.02	0.94 ± 0.01
	k-NN	0.98 ± 0.01	0.94 ± 0.01	0.98 ± 0.01
	Logistic regression	0.97 ± 0.01	0.93 ± 0.01	0.96 ± 0.01
	Random forest	0.97 ± 0.01	0.90 ± 0.01	0.97 ± 0.01
	Super learner	0.98 ± 0.01	0.88 ± 0.02	0.97 ± 0.01
	SVM	0.97 ± 0.01	0.89 ± 0.02	0.97 ± 0.01
GR				
	AdaBoost classifier	0.98 ± 0.01	0.93 ± 0.01	0.96 ± 0.01
	Bagging classifier	0.98 ± 0.01	0.94 ± 0.01	0.95 ± 0.01
	Decision tree classifier	0.97 ± 0.01	0.93 ± 0.01	0.95 ± 0.01
	Gaussian naive Bayes	0.95 ± 0.01	0.92 ± 0.01	0.93 ± 0.01
	k-NN	0.98 ± 0.01	0.94 ± 0.01	0.97 ± 0.01
	Logistic regression	0.98 ± 0.01	0.92 ± 0.01	0.96 ± 0.01
	Random forest	0.97 ± 0.01	0.92 ± 0.01	0.97 ± 0.01
	Super learner	0.98 ± 0.01	0.93 ± 0.01	0.98 ± 0.01
	SVM	0.98 ± 0.01	0.93 ± 0.01	0.98 ± 0.01
PR				
	AdaBoost classifier	0.98 ± 0.01	0.87 ± 0.01	0.98 ± 0.01
	Bagging classifier	0.99 ± 0.00	0.86 ± 0.01	0.95 ± 0.01
	Decision tree classifier	0.98 ± 0.01	0.86 ± 0.02	0.96 ± 0.01
	Gaussian naive Bayes	0.98 ± 0.01	0.80 ± 0.02	0.93 ± 0.01
	k-NN	0.98 ± 0.01	0.87 ± 0.01	0.99 ± 0.00
	Logistic regression	0.99 ± 0.00	0.85 ± 0.02	0.98 ± 0.01
	Random forest	0.98 ± 0.01	0.85 ± 0.02	0.97 ± 0.01
	Super learner	0.98 ± 0.01	0.86 ± 0.02	0.99 ± 0.00
	SVM	0.98 ± 0.01	0.87 ± 0.02	0.99 ± 0.00

± = SD.

Table 4. Average accuracy of different algorithms for three class approach for six receptors using MACCS fingerprints as input features on the repeated k-fold CV

Receptor	Algorithm	Agonist Accuracy	Antagonist Accuracy	Binders Accuracy
AR	AdaBoost classifier	0.93 ± 0.01	0.88 ± 0.01	0.92 ± 0.01
	Bagging classifier	0.94 ± 0.01	0.88 ± 0.01	0.93 ± 0.01
	Decision tree classifier	0.91 ± 0.01	0.84 ± 0.02	0.92 ± 0.01
	Gaussian naive Bayes	0.91 ± 0.01	0.78 ± 0.02	0.83 ± 0.02
	k-NN	0.96 ± 0.01	0.87 ± 0.01	0.98 ± 0.01
	Logistic regression	0.85 ± 0.02	0.82 ± 0.02	0.91 ± 0.01
	Random forest	0.94 ± 0.01	0.88 ± 0.01	0.95 ± 0.01
	Super learner	0.95 ± 0.01	0.87 ± 0.02	0.98 ± 0.01
	SVM	0.96 ± 0.01	0.89 ± 0.01	0.97 ± 0.01
ERA				
	AdaBoost classifier	0.87 ± 0.01	0.90 ± 0.01	0.92 ± 0.01
	Bagging classifier	0.88 ± 0.01	0.91 ± 0.01	0.93 ± 0.01
	Decision tree classifier	0.81 ± 0.02	0.86 ± 0.01	0.91 ± 0.01
	Gaussian naive Bayes	0.73 ± 0.02	0.76 ± 0.02	0.83 ± 0.02
	k-NN	0.86 ± 0.01	0.89 ± 0.01	0.95 ± 0.01
	Logistic regression	0.74 ± 0.02	0.79 ± 0.02	0.90 ± 0.01
	Random forest	0.87 ± 0.01	0.89 ± 0.01	0.94 ± 0.01
	Super learner	0.83 ± 0.02	0.83 ± 0.02	0.95 ± 0.01
	SVM	0.85 ± 0.01	0.91 ± 0.01	0.95 ± 0.01
ERB				
	AdaBoost classifier	0.96 ± 0.01	0.93 ± 0.01	0.95 ± 0.01
	Bagging classifier	0.96 ± 0.00	0.94 ± 0.01	0.96 ± 0.01
	Decision tree classifier	0.95 ± 0.01	0.89 ± 0.01	0.94 ± 0.01
	Gaussian naive Bayes	0.95 ± 0.01	0.79 ± 0.02	0.86 ± 0.01
	k-NN	0.95 ± 0.01	0.91 ± 0.01	0.98 ± 0.01
	Logistic regression	0.90 ± 0.01	0.78 ± 0.02	0.94 ± 0.01
	Random forest	0.97 ± 0.01	0.93 ± 0.01	0.97 ± 0.01
	Super learner	0.96 ± 0.01	0.85 ± 0.02	0.97 ± 0.01
	SVM	0.98 ± 0.01	0.84 ± 0.02	0.98 ± 0.01
FXR				
	AdaBoost classifier	0.96 ± 0.01	0.94 ± 0.01	0.95 ± 0.01
	Bagging classifier	0.97 ± 0.01	0.96 ± 0.00	0.96 ± 0.01
	Decision tree classifier	0.95 ± 0.01	0.92 ± 0.01	0.93 ± 0.01
	Gaussian naive Bayes	0.87 ± 0.02	0.79 ± 0.02	0.84 ± 0.02
	k-NN	0.97 ± 0.01	0.94 ± 0.01	0.97 ± 0.01
	Logistic regression	0.92 ± 0.01	0.81 ± 0.05	0.91 ± 0.01
	Random forest	0.97 ± 0.01	0.93 ± 0.01	0.97 ± 0.01
	Super learner	0.96 ± 0.01	0.81 ± 0.04	0.96 ± 0.01
	SVM	0.98 ± 0.01	0.78 ± 0.02	0.97 ± 0.01
GR				
	AdaBoost classifier	0.96 ± 0.01	0.92 ± 0.01	0.93 ± 0.01
	Bagging classifier	0.97 ± 0.01	0.93 ± 0.01	0.94 ± 0.01
	Decision tree classifier	0.96 ± 0.01	0.89 ± 0.01	0.92 ± 0.01
	Gaussian naive Bayes	0.85 ± 0.02	0.83 ± 0.02	0.86 ± 0.01
	k-NN	0.98 ± 0.01	0.94 ± 0.01	0.98 ± 0.01
	Logistic regression	0.92 ± 0.01	0.87 ± 0.02	0.92 ± 0.01
	Random forest	0.97 ± 0.01	0.93 ± 0.01	0.95 ± 0.01
	Super learner	0.85 ± 0.02	0.83 ± 0.02	0.86 ± 0.01
	SVM	0.98 ± 0.01	0.94 ± 0.01	0.97 ± 0.01
PPARG				
	AdaBoost classifier	0.93 ± 0.01	0.94 ± 0.01	0.92 ± 0.01
	Bagging classifier	0.94 ± 0.01	0.95 ± 0.00	0.94 ± 0.01
	Decision tree classifier	0.92 ± 0.01	0.91 ± 0.01	0.91 ± 0.01
	Gaussian naive Bayes	0.86 ± 0.01	0.92 ± 0.01	0.84 ± 0.01
	k-NN	0.97 ± 0.01	0.92 ± 0.01	0.96 ± 0.01
	Logistic regression	0.91 ± 0.01	0.75 ± 0.02	0.90 ± 0.01
	Random forest	0.95 ± 0.01	0.95 ± 0.01	0.95 ± 0.01
	Super learner	0.86 ± 0.01	0.92 ± 0.01	0.84 ± 0.01
	SVM	0.96 ± 0.01	0.82 ± 0.02	0.96 ± 0.01
PR				
	AdaBoost classifier	0.98 ± 0.01	0.86 ± 0.01	0.95 ± 0.01
	Bagging classifier	0.98 ± 0.01	0.85 ± 0.01	0.94 ± 0.01
	Decision tree classifier	0.97 ± 0.01	0.82 ± 0.02	0.93 ± 0.01

(*Continued*)

Table 4. Continued

Receptor	Algorithm	Agonist Accuracy	Antagonist Accuracy	Binders Accuracy
	Gaussian naive Bayes	0.96 ± 0.01	0.77 ± 0.02	0.85 ± 0.02
	k-NN	0.98 ± 0.01	0.86 ± 0.02	0.99 ± 0.00
	Logistic regression	0.95 ± 0.01	0.81 ± 0.02	0.95 ± 0.01
	Random forest	0.99 ± 0.00	0.86 ± 0.02	0.96 ± 0.01
	Super learner	0.98 ± 0.01	0.86 ± 0.02	0.98 ± 0.01
	SVM	0.98 ± 0.01	0.88 ± 0.01	0.98 ± 0.01

± = SD.

Table 5. Comparison of the performance of different classifiers on the validation set for three class approach for AR using ECFP4 fingerprints as input features

Binding class	Methods	BA	Sn	Sp	MCC	PR AUC	TP	TN	FP	FN
Agonist	AdaBoost classifier	0.85	0.74	0.96	0.68	0.73	87	1084	30	42
	Bagging classifier	0.85	0.71	1.00	0.81	0.79	83	1123	34	3
	Decision tree classifier	0.86	0.74	0.98	0.73	0.73	87	1100	30	26
	Gaussian naive Bayes	0.85	0.75	0.94	0.61	0.69	88	1059	29	67
	k-NN	0.86	0.77	0.95	0.65	0.81	90	1069	27	57
	Logistic regression	0.85	0.78	0.93	0.59	0.73	91	1044	26	82
	Random forest	0.86	0.75	0.98	0.74	0.80	88	1101	29	25
	Super learner	0.87	0.77	0.98	0.75	0.81	90	1099	27	27
	SVM	0.87	0.77	0.97	0.74	0.80	90	1097	27	29
Antagonist										
	AdaBoost classifier	0.79	0.63	0.95	0.62	0.78	153	928	88	53
	Bagging classifier	0.72	0.44	1.00	0.61	0.78	105	978	136	3
	Decision tree classifier	0.75	0.55	0.95	0.56	0.65	133	930	108	51
	Gaussian naive Bayes	0.80	0.74	0.86	0.55	0.69	178	846	63	135
	k-NN	0.80	0.68	0.92	0.61	0.80	165	906	76	75
	Logistic regression	0.81	0.73	0.89	0.59	0.76	176	874	65	107
	Random forest	0.83	0.76	0.90	0.62	0.82	183	881	58	100
	Super learner	0.83	0.72	0.94	0.67	0.83	173	922	68	59
	SVM	0.84	0.76	0.92	0.66	0.84	183	901	58	80
Binder										
	AdaBoost classifier	0.94	0.91	0.98	0.88	0.96	254	1026	26	25
	Bagging classifier	0.90	0.81	0.99	0.86	0.94	226	1044	54	7
	Decision tree classifier	0.93	0.90	0.97	0.86	0.89	251	1018	29	33
	Gaussian naive Bayes	0.92	0.94	0.90	0.77	0.85	263	949	17	102
	k-NN	0.96	0.94	0.99	0.94	0.97	262	1041	18	10
	Logistic regression	0.96	0.94	0.97	0.89	0.97	264	1018	16	33
	Random forest	0.94	0.90	0.97	0.87	0.97	253	1020	27	31
	Super learner	0.97	0.95	0.99	0.95	0.98	265	1043	15	8
	SVM	0.97	0.94	0.99	0.94	0.98	264	1042	16	9

[a]BA, balanced accuracy; Sn, sensitivity; Sp, specificity; MCC, Mathew correlation coefficient; PR AUC, precision-recall curve; TP, true positive; TN, true negative; FN, false negative; FP, false positive.

online at *Briefings in Bioinformatics*) and 84 and 85% for MACCS fingerprints (Table S8**,** see Supplementary Data available online at *Briefings in Bioinformatics*). The model developed using SVM combined with ECFP4 fingerprints had a maximum MCC value of 0.82 with the specificity, sensitivity and balanced accuracy of 94, 94 and 89%, respectively. Similar performance has been observed for other classifiers with ECFP4 and MACCS fingerprints.

Models for FXR and PPARG

Binding class models for FXR and PPARG

Average accuracy for repeated stratified k-fold CV for the classifiers based on the ECFP4 and MACCS fingerprints for different classes of FXR and PPARG are given in Tables 3 and 4**,** respectively, demonstrating that all classifiers have achieved accuracies of >90% at identifying FXR agonist and binders. Specifically, Bagging classifier, k-NN, Random forest, Super learner and SVM classifiers attained an accuracy of >95% at identifying FXR agonist and binders with MACCS fingerprints. For the FXR antagonist dataset, AdaBoost classifier, Bagging classifier, decision tree and random forest have accuracies of >90% with repeated stratified k-fold CV with ECFP4 and MACCS. The performance of different classifiers for different classes of FXR and PPARG on the validation dataset are given in Table S9 (see Supplementary Data available online at *Briefings in Bioinformatics*) (ECFP4)**,** S10 (MACCS), Table S11 (see

Table 6. Comparison of the performance of different classifiers on the validation set for three class approach for AR using MACCS fingerprints as input features

Binding class	Method	BA	Sn	Sp	MCC	PR AUC	TP	TN	FN	FP
Agonist	AdaBoost classifier	0.86	0.73	0.98	0.76	0.77	85	1109	32	17
	Bagging classifier	0.82	0.64	1.00	0.76	0.78	75	1121	42	5
	Decision tree classifier	0.84	0.72	0.97	0.69	0.68	84	1093	33	33
	Gaussian naive Bayes	0.80	0.78	0.83	0.42	0.53	91	930	26	196
	k-NN	0.86	0.74	0.98	0.73	0.78	87	1100	30	26
	Logistic regression	0.82	0.77	0.87	0.47	0.75	90	975	27	151
	Random forest	0.86	0.75	0.96	0.69	0.79	88	1085	29	41
	Super learner	0.87	0.79	0.96	0.71	0.81	92	1084	25	42
	SVM	0.86	0.79	0.94	0.64	0.79	92	1060	25	66
Antagonist										
	AdaBoost classifier	0.77	0.58	0.96	0.62	0.78	139	946	102	35
	Bagging classifier	0.73	0.46	0.99	0.61	0.80	112	974	129	7
	Decision tree classifier	0.79	0.63	0.94	0.60	0.71	152	922	89	59
	Gaussian naive Bayes	0.76	0.78	0.74	0.43	0.62	188	727	53	254
	k-NN	0.82	0.76	0.88	0.60	0.79	184	866	57	115
	Logistic regression	0.82	0.79	0.85	0.56	0.78	190	829	51	152
	Random forest	0.84	0.77	0.91	0.65	0.84	185	894	56	87
	Super learner	0.84	0.75	0.93	0.66	0.82	180	910	61	71
	SVM	0.84	0.80	0.89	0.64	0.83	192	874	49	107
Binder										
	AdaBoost classifier	0.92	0.88	0.97	0.85	0.95	247	1016	33	35
	Bagging classifier	0.91	0.84	0.98	0.85	0.95	234	1032	46	19
	Decision tree classifier	0.95	0.93	0.97	0.89	0.93	259	1023	21	28
	Gaussian naive Bayes	0.85	0.93	0.78	0.59	0.79	259	816	21	235
	k-NN	0.95	0.93	0.97	0.88	0.96	259	1018	21	33
	Logistic regression	0.94	0.94	0.94	0.83	0.94	263	988	17	63
	Random forest	0.95	0.94	0.96	0.87	0.97	262	1010	18	41
	Super learner	0.96	0.94	0.98	0.92	0.97	264	1033	16	18
	SVM	0.96	0.94	0.98	0.92	0.97	262	1034	18	17

[a]BA, balanced accuracy; Sn, sensitivity; Sp, specificity; MCC, Mathew correlation coefficient; PR AUC, precision-recall curve; TP, true positive; TN, true negative; FN, false negative; FP, false positive.

Table 7. Comparison of the performance of different classifiers on the validation set for AR effector dataset

Finger-print	Method	BA	Sn	Sp	MCC	PR AUC	TP	TN	FN	FP
ECFP4	AdaBoost classifier	0.83	0.76	0.91	0.67	0.86	331	949	104	96
	Bagging classifier	0.81	0.65	0.98	0.71	0.86	281	1024	154	21
	Decision tree classifier	0.80	0.64	0.96	0.67	0.77	280	1005	155	40
	Gaussian naive Bayes	0.83	0.72	0.94	0.69	0.82	312	981	123	64
	k-NN	0.85	0.73	0.98	0.77	0.89	316	1026	119	19
	Logistic regression	0.83	0.79	0.86	0.64	0.83	344	903	91	142
	Random forest	0.86	0.78	0.93	0.73	0.89	341	975	94	70
	Super learner	0.86	0.76	0.96	0.75	0.89	332	1000	103	45
	SVM	0.87	0.76	0.98	0.78	0.90	329	1019	106	26
MACSS										
	AdaBoost classifier	0.82	0.72	0.93	0.67	0.86	315	967	120	78
	Bagging classifier	0.82	0.68	0.96	0.70	0.87	297	1007	138	38
	Decision tree classifier	0.83	0.73	0.93	0.68	0.77	318	971	117	74
	Gaussian naive Bayes	0.78	0.78	0.78	0.53	0.69	341	816	94	229
	k-NN	0.86	0.78	0.94	0.74	0.89	338	986	97	59
	Logistic regression	0.84	0.83	0.85	0.65	0.86	359	888	76	157
	Random forest	0.85	0.79	0.91	0.70	0.90	344	954	91	91
	Super learner	0.86	0.79	0.93	0.73	0.89	345	969	90	76
	SVM	0.87	0.79	0.94	0.75	0.90	345	982	90	63

[a]BA, balanced accuracy; Sn, sensitivity; Sp, specificity; MCC, Mathew correlation coefficient; PR AUC, Precision-recall curve; TP, true positive; TN, true negative; FN, false negative; FP, false positive.

Table 8. Average accuracy of different algorithms for effector dataset of different receptors using ECFP4 fingerprints and MACCS fingerprints as input features on the repeated k-fold CV

		Accuracy								
Finger-print	**Methods**	**AR**	**ERA**	**ERB**	**FXR**	**GR**	**PPARD**	**PPARG**	**PR**	**RXR**
ECFP4	AdaBoost classifier	0.87 ± 0.01	0.83 ± 0.01	0.92 ± 0.01	0.93 ± 0.01	0.92 ± 0.01	0.97 ± 0.01	0.92 ± 0.01	0.87 ± 0.01	0.95 ± 0.01
	Bagging classifier	0.87 ± 0.01	0.85 ± 0.01	0.93 ± 0.01	0.94 ± 0.01	0.93 ± 0.01	0.98 ± 0.01	0.92 ± 0.01	0.87 ± 0.01	0.96 ± 0.01
	Decision tree classifier	0.87 ± 0.01	0.84 ± 0.01	0.93 ± 0.01	0.93 ± 0.01	0.92 ± 0.01	0.97 ± 0.01	0.91 ± 0.01	0.87 ± 0.01	0.95 ± 0.01
	Gaussian naive Bayes	0.86 ± 0.01	0.82 ± 0.01	0.90 ± 0.01	0.94 ± 0.01	0.91 ± 0.01	0.98 ± 0.01	0.91 ± 0.01	0.85 ± 0.01	0.96 ± 0.01
	k-NN	0.90 ± 0.01	0.85 ± 0.01	0.94 ± 0.01	0.95 ± 0.01	0.95 ± 0.01	0.98 ± 0.01	0.94 ± 0.01	0.90 ± 0.01	0.96 ± 0.01
	Logistic regression	0.84 ± 0.01	0.79 ± 0.02	0.90 ± 0.01	0.91 ± 0.01	0.91 ± 0.01	0.97 ± 0.01	0.91 ± 0.01	0.85 ± 0.01	0.93 ± 0.01
	Random forest	0.88 ± 0.01	0.84 ± 0.01	0.94 ± 0.01	0.94 ± 0.01	0.94 ± 0.01	0.97 ± 0.01	0.93 ± 0.01	0.87 ± 0.01	0.96 ± 0.01
	Super learner	0.89 ± 0.01	0.85 ± 0.01	0.94 ± 0.01	0.94 ± 0.01	0.95 ± 0.01	0.98 ± 0.01	0.94 ± 0.01	0.89 ± 0.01	0.96 ± 0.01
	SVM	0.90 ± 0.01	0.86 ± 0.01	0.94 ± 0.01	0.95 ± 0.01	0.95 ± 0.01	0.98 ± 0.01	0.94 ± 0.01	0.90 ± 0.01	0.96 ± 0.01
MACCS										
	AdaBoost classifier	0.87 ± 0.01	0.83 ± 0.01	0.92 ± 0.01	0.93 ± 0.01	0.92 ± 0.01	0.97 ± 0.01	0.91 ± 0.01	0.88 ± 0.01	0.95 ± 0.01
	Bagging classifier	0.88 ± 0.01	0.85 ± 0.01	0.93 ± 0.01	0.94 ± 0.01	0.93 ± 0.01	0.97 ± 0.01	0.92 ± 0.01	0.88 ± 0.01	0.96 ± 0.01
	Decision tree classifier	0.87 ± 0.01	0.84 ± 0.01	0.92 ± 0.01	0.93 ± 0.01	0.93 ± 0.01	0.97 ± 0.01	0.92 ± 0.01	0.87 ± 0.01	0.96 ± 0.01
	Gaussian naive Bayes	0.88 ± 0.01	0.84 ± 0.01	0.93 ± 0.01	0.92 ± 0.01	0.94 ± 0.01	0.97 ± 0.01	0.93 ± 0.01	0.89 ± 0.01	0.96 ± 0.01
	k-NN	0.89 ± 0.01	0.84 ± 0.01	0.93 ± 0.01	0.94 ± 0.01	0.94 ± 0.01	0.97 ± 0.01	0.93 ± 0.01	0.89 ± 0.01	0.96 ± 0.01
	Logistic regression	0.84 ± 0.01	0.80 ± 0.02	0.88 ± 0.01	0.87 ± 0.01	0.91 ± 0.01	0.94 ± 0.01	0.90 ± 0.01	0.86 ± 0.01	0.93 ± 0.01
	Random forest	0.87 ± 0.01	0.84 ± 0.01	0.92 ± 0.01	0.92 ± 0.01	0.93 ± 0.01	0.97 ± 0.01	0.92 ± 0.01	0.87 ± 0.01	0.96 ± 0.01
	Super learner	0.88 ± 0.01	0.84 ± 0.01	0.93 ± 0.01	0.92 ± 0.01	0.94 ± 0.01	0.97 ± 0.01	0.93 ± 0.01	0.89 ± 0.01	0.96 ± 0.01
	SVM	0.89 ± 0.01	0.85 ± 0.01	0.93 ± 0.01	0.94 ± 0.01	0.95 ± 0.01	0.98 ± 0.01	0.94 ± 0.01	0.90 ± 0.01	0.96 ± 0.01

± = SD.

Supplementary Data available online at *Briefings in Bioinformatics*) (ECFP4) and S12 (MACCS), respectively. The results demonstrate that Super learner has attained better performance for agonists and binders of FXR with different fingerprints. Similar performance has been achieved for PPARG agonists and binders. Poor performance of antagonist models was obtained on the validation set for all the classifiers for both FXR and PPARG due to the sample size of the training dataset.

Models for GR and PR

Binding class models for GR and PR

Average repeated stratified k-fold CV accuracy for the classifiers based on the ECFP4 and MACCS fingerprints and for different classes of GR and PR is given in Tables 3 and 4, respectively. Results show that SVM and Super learner algorithms have higher accuracy in identifying agonists and binders for GR and PR based on repeated stratified k-fold CV. The performance of different classifiers for different classes of GR and PR on the validation dataset are given in Tables S13 (see Supplementary Data available online at *Briefings in Bioinformatics*) (ECFP4), S14 (MACCS), Table S15 (see Supplementary Data available online at *Briefings in Bioinformatics*) (ECFP4) and S16 (MACCS), respectively. Results show that random forest, Super learner and SVM have good performance scores for the three classes of GR and PR with different features.

Effector models for FXR, GR and PR, PPARG, PPARD and RXR

Data availability for antagonists of PPARD and RXR is limited; hence we have not modeled the different classes. We merged the dataset as described in the materials and methods to create an effector dataset for these receptors. Performance measures on the repeated stratified k-fold CV for FXR, GR, PR, PPARG, PPARD and RXR are given in Tables 3 and 4 for ECFP4 and MACCS fingerprints, respectively. Results show high accuracy across these NRs for all classifiers with both fingerprint types. The different performance measures on the datasets for FXR, GR, PR, PPARD, PPARG and RXR are given in Tables S17–S22 (see Supplementary Data available online at *Briefings in Bioinformatics*), respectively. Tables 3 and 4 show that Super learner and SVM have both attained accuracies of >90% for the effector dataset of these receptors. Tables S17 and S19 (see Supplementary Data available online at *Briefings in Bioinformatics*) for FXR and PR show that most of the classifiers attained high accuracy for both fingerprint types. The random forest, k-NN and SVM classifiers with ECFP4 fingerprints showed similar sensitivity/specificity of 94 and 95%/94 and 95%, respectively, with an MCC value of 0.75–0.76 on the validation dataset. The results for the ligand-binding predictions for GR, PPARD, PPARG and RXR (Tables S18, S20, S21 and S22**,** see Supplementary Data available online at *Briefings in Bioinformatics*) show that the SVM-based models achieved slightly higher accuracy and MCC score than other evaluated algorithms.

Applicability domain on the validation set

We applied the applicability domain to the validation set and removed the unreliable data points that were thus identified. Then we evaluated the performance of the SVM and Super learner models on the remaining reliable

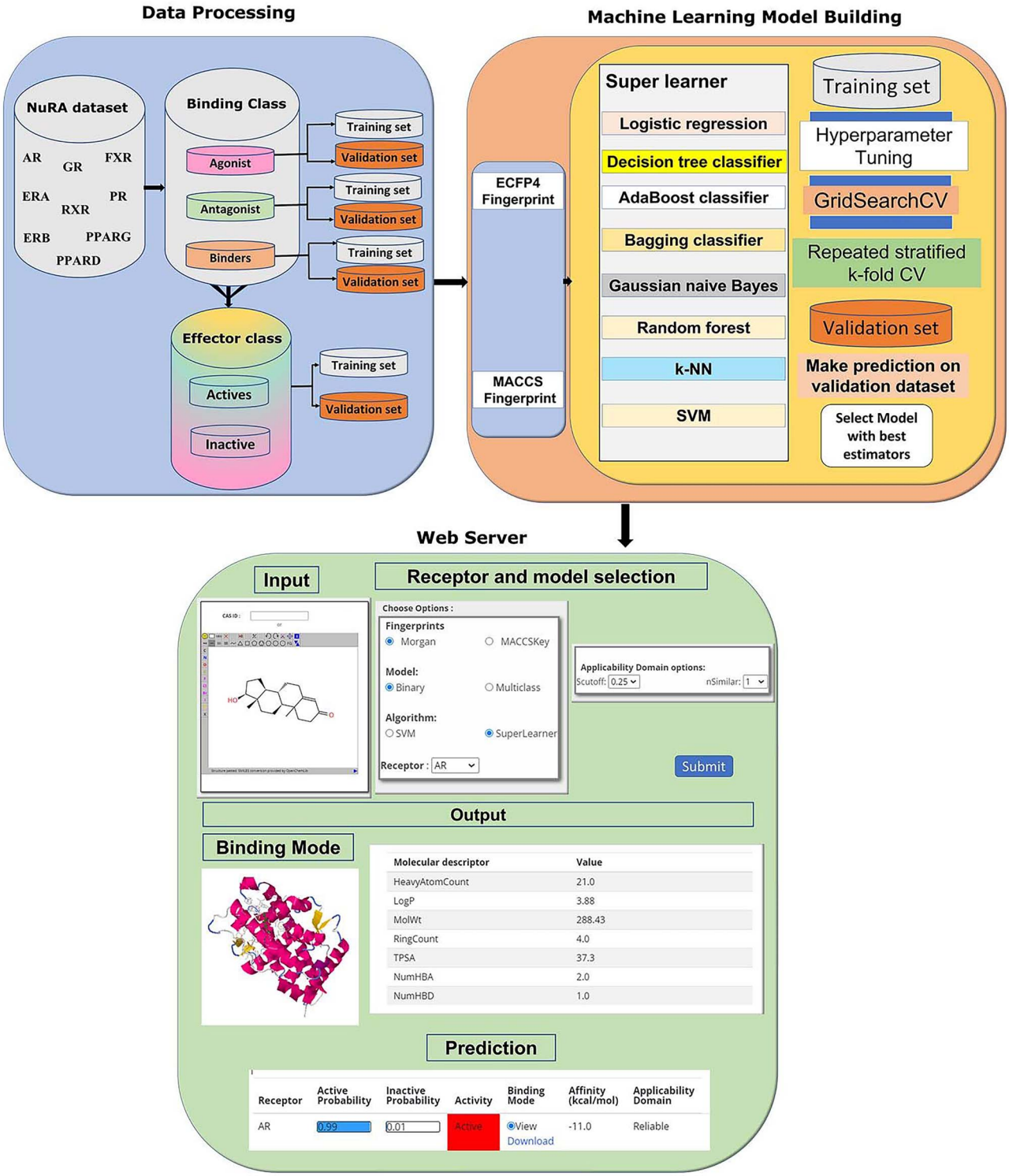

Figure 1. Schematic of the development and use of NR-Toxpred webserver.

data points from the validation dataset. The results on the validation dataset after filtering the dataset through the applicability domain for the reliability of the prediction are given in supporting information as a Microsoft Excel workbook (S23–S81). The results show that including the applicability domain with SVM and Super learner models with ECFP4 fingerprints improves the model's performance.

Implementation of NR-ToxPred web server

Based on our trained and validated best-performing models, we have developed a web-based application named NR-ToxPred with a user-friendly interface to assist the scientific community (Figure 1). We implemented the best SVM and Super learner-based model for all nine NRs on the webserver. The user interface of the NR-ToxPred server allows for different formats to

submit small molecules. Users can sketch the structure using a simple drawing interface, give SMILES codes as text input in the drawing interface or input CAS ID data as the search criteria. Users can upload a two-column file with SMILES codes and corresponding names in a comma-separated CSV format for multiple ligand predictions. For the single structure input, in addition to the tabulated results for each receptor, if the chemical is a predicted ligand, it is subsequently docked to the matching receptor(s). Users can select the applicability domain criteria (S_{cutoff} and N_{min}). The NR-ToxPred web service can be accessed at http://nr-toxpred.cchem.berkeley.edu/.

Limitations of the models

In this study, we developed different ML models for predicting agonist, antagonist, binders binding class (each binding class as binary: active versus inactive) and also effectors (binding versus nonbinding). Then, as needed, we constrained these to the applicability domain within each receptor according to the available number of chemicals in each class in the dataset. We initially found poor predictive power for the antagonist models of FXR, but this was overcome by setting stricter criteria for the applicability domain. For PPARG, PPARD and RXR models, we collapsed the agonist and antagonist from the dataset into one category, also known as effector, due to the limitations in the available number of chemicals in each antagonist category in the dataset. The models herein are thus limited to predicting only the binding of the small molecules to these NRs. They are not capable of distinguishing agonists versus antagonists. However, this distinction is easily determined in an experimental setting once the binding candidates are identified. This experimental testing is much more tractable with the computationally shortlisted dataset than testing the whole set of chemicals. For the other NRs, our predictions are well-validated and robust with more robust data.

Conclusions

ML-based classification models for nine different human NRs have been developed and validated with the NuRA dataset. These models were trained using fingerprints as features to predict the binding class (agonist, antagonist, binders) for active and inactive chemicals binding to these selected receptors as well as effectors class (binding versus nonbinding). Furthermore, we show that the applicability domain and chemical space variation of actives and inactives are essential for accurate binding predictions. We developed a publicly available web-based application named NR-ToxPred with a user-friendly interface to assist the scientific community using these tools and approaches.

Key Points

- NRs are important targets for the toxic effects of many xenobiotic chemicals that can act as agonists or antagonists to the receptor in question.
- Models have been trained on the NuRA dataset with different ML algorithms for developing reliable models.
- NR-ToxPred is a publicly available web server for predicting the binding of chemical molecules to multiple different NRs.
- The web server can be accessed at http://nr-toxpred.cchem.berkeley.edu/

Supplementary data

Supplementary data are available at *Briefings in Bioinformatics* online.

Acknowledgments

The authors would like to thank Dr Zhongyu Wang of the Key Laboratory of Industrial Ecology and Environmental Engineering (MOE), School of Environmental Science and Technology, Dalian University of Technology, Dalian 116024, China, who provided the script for calculating the applicability domain. ERA would like to dedicate this work to Dr V Subramanian.

Funding

Office of Environmental Health Hazard Assessment of the California Environmental Protection Agency (EPA) (USA) (17-0023, 17-E0024); USDA National Institute of Food and Agriculture, Hatch project (1002182); National Institutes of Health (S10OD023532 to K.A.D.).

References

1. Weikum ER, Liu X, Ortlund EA. The nuclear receptor superfamily: a structural perspective. *Protein Sci* 2018;**27**:1876–92.
2. Gronemeyer H, Gustafsson JA, Laudet V. Principles for modulation of the nuclear receptor superfamily. *Nat Rev Drug Discov* 2004;**3**:950–64.
3. Germain P, Staels B, Dacquet C, *et al.* Overview of nomenclature of nuclear receptors. *Pharmacol Rev* 2006;**58**:685–704.
4. Beato M, Klug J. Steroid hormone receptors: an update. *Hum Reprod Update* 2000;**6**:225–36.
5. Sonoda J, Pei L, Evans RM. Nuclear receptors: decoding metabolic disease. *FEBS Lett* 2008;**582**:2–9.
6. Zhao L, Zhou S, Gustafsson JA. Nuclear receptors: recent drug discovery for cancer therapies. *Endocr Rev* 2019;**40**:1207–49.
7. Smith MT, Guyton KZ, Gibbons CF, *et al.* Key characteristics of carcinogens as a basis for organizing data on mechanisms of carcinogenesis. *Environ Health Perspect* 2016;**124**:713–21.
8. Smith MT, Guyton KZ, Kleinstreuer N, *et al.* The key characteristics of carcinogens: relationship to the hallmarks of cancer, relevant biomarkers, and assays to measure them. *Cancer Epidemiol Biomarkers Prev* 2020;**29**:1887–903.
9. Wang Z, Walker GW, Muir DCG, *et al.* Toward a global understanding of chemical pollution: a first comprehensive analysis

of national and regional chemical inventories. *Environ Sci Technol* 2020;**54**:2575–84.
10. Mansouri K, Abdelaziz A, Rybacka A, *et al.* CERAPP: collaborative estrogen receptor activity prediction project. *Environ Health Perspect* 2016;**124**:1023–33.
11. Mansouri K, Kleinstreuer N, Abdelaziz AM, *et al.* CoMPARA: collaborative modeling project for androgen receptor activity. *Environ Health Perspect* 2020;**128**:27002.
12. Judson RS, Friedman KP, Houck K, *et al.* New approach methods for testing chemicals for endocrine disruption potential. *Curr Opin Toxicol* 2018;**9**:40–7.
13. Todorov M. Computational tools for prediction of nuclear receptor mediated effects. *East Acad J* 2015;**1**:46–56.
14. Piir G, Sild S, Maran U. Binary and multi-class classification for androgen receptor agonists, antagonists and binders. *Chemosphere* 2021;**262**:128313.
15. Singam ERA, Tachachartvanich P, Fourches D, *et al.* Structure-based virtual screening of perfluoroalkyl and polyfluoroalkyl substances (PFASs) as endocrine disruptors of androgen receptor activity using molecular docking and machine learning. *Environ Res* 2020;**190**:109920.
16. Azhagiya Singam ER, Tachachartvanich P, La Merrill MA, *et al.* Structural dynamics of agonist and antagonist binding to the androgen receptor. *J Phys Chem B* 2019;**123**:7657–66.
17. Trisciuzzi D, Alberga D, Mansouri K, *et al.* Predictive structure-based toxicology approaches to assess the androgenic potential of chemicals. *J Chem Inf Model* 2017;**57**:2874–84.
18. Ekins S, Kortagere S, Iyer M, *et al.* Challenges predicting ligand-receptor interactions of promiscuous proteins: the nuclear receptor PXR. *PLoS Comput Biol* 2009;**5**:e1000594.
19. Cotterill J, Palazzolo L, Ridgway C, *et al.* Predicting estrogen receptor binding of chemicals using a suite of in silico methods–complementary approaches of (Q) SAR, molecular docking and molecular dynamics. *Toxicol Appl Pharmacol* 2019;**378**:114630.
20. Sun L, Yang H, Cai Y, *et al.* In silico prediction of endocrine disrupting chemicals using single-label and multilabel models. *J Chem Inf Model* 2019;**59**:973–82.
21. Capuzzi SJ, Politi R, Isayev O, *et al.* QSAR modeling of Tox21 challenge stress response and nuclear receptor signaling toxicity assays. *Front Environ Sci* 2016;**4**:3.
22. Jaladanki CK, He Y, Zhao LN, *et al.* Virtual screening of potentially endocrine-disrupting chemicals against nuclear receptors and its application to identify PPARγ-bound fatty acids. *Arch Toxicol* 2021;**95**:355–74.
23. Valsecchi C, Grisoni F, Motta S, *et al.* NURA: a curated dataset of nuclear receptor modulators. *Toxicol Appl Pharmacol* 2020;**407**:115244.
24. Tice RR, Austin CP, Kavlock RJ, *et al.* Improving the human hazard characterization of chemicals: a Tox21 update. *Environ Health Perspect* 2013;**121**:756–65.
25. Gaulton A, Hersey A, Nowotka M, *et al.* The ChEMBL database in 2017. *Nucleic Acids Res* 2017;**45**:D945–54.
26. Liu T, Lin Y, Wen X, *et al.* BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities. *Nucleic Acid Res* 2007;**35**:D198–201.
27. Réau M, Lagarde N, JFO Z, *et al.* Nuclear receptors database including negative data (NR-DBIND): a database dedicated to nuclear receptors binding data including negative data and pharmacological profile: Miniperspective. *J Med Chem* 2018;**62**:2894–904.
28. Valsecchi C, Collarile M, Grisoni F, *et al.* Predicting molecular activity on nuclear receptors by multitask neural networks. *J Chemometr* 2020;**36**:e3325.
29. Michael R, Berthold N, Dill F, *et al.* KNIME: the Konstanz Information Miner. In: *Data Analysis, Machine Learning and Applications Studies in Classification, Data Analysis, and Knowledge Organization*. Berlin.[Google Scholar]: Springer, 2008.
30. Landrum G. *RDKit: A Software Suite for Cheminformatics, Computational Chemistry, and Predictive Modeling*. Cambridge: Academic Press, 2013.
31. Van der Laan MJ, Polley EC, Hubbard AE. Super learner. *Statistical applications in genetics and molecular biology*, 2007;6(1).
32. Freund Y, Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. *J Comp Syst Sci* 1997;**55**:119–39.
33. Wright RE. *Logistic Regression*. In L. G. Grimm & P. R. Yarnold (Eds.), *Reading and understanding multivariate statistics* (pp. 217–244). American Psychological Association. 1995.
34. Breiman L. Random forests. *Mach Learn* 2001;**45**:5–32.
35. Wang L. *Support Vector Machines: Theory and Applications*. Springer Science & Business Media, Berlin, Springer 2005 pp. 365–386.
36. Altman NS. An introduction to kernel and nearest-neighbor nonparametric regression. *Am Stat* 1992;**46**:175–85.
37. Breiman L. Bagging predictors. *Mach Learn* 1996;**24**:123–40.
38. Chan TF, Golub GH, LeVeque RJ. Updating formulae and a pairwise algorithm for computing sample variances. In: *COMPSTAT 1982 5th Symposium held at Toulouse 1982*. Physica, Heidelberg, CA, USA, 1982, 30–41.
39. Loh WY. Classification and regression trees. *Wiley Interdiscip Rev Data Min Knowl Discov* 2011;**1**:14–23.
40. Pedregosa F, Varoquaux G, Gramfort A, *et al.* Scikit-learn: machine learning in python. *J Mach Learn Res* 2011;**12**:2825–30.
41. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme, Biochimica et Biophysica Acta (BBA)-protein. *Structure* 1975;**405**:442–51.
42. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One* 2015;**10**:e0118432.
43. Mosley L. *A Balanced Approach to the Multi-Class Imbalance Problem*. 2013. (Ph.D. thesis) Iowa State University Capstones.
44. Wang Z, Chen J, Hong H. Applicability domains enhance application of PPAR gamma agonist classifiers trained by drug-like compounds to environmental chemicals. *Chem Res Toxicol* 2020;**33**:1382–8.
45. Herraez A. Biomolecules in the computer: Jmol to the rescue. *Biochem Mol Biol Educ* 2006;**34**:255–61.
46. Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem* 2010;**31**:455–61.
47. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* 2020;**21**:1–13.