Published in final edited form as:

Mol Pharm. 2019 June 03; 16(6): 2605–2615. doi:10.1021/acs.molpharmaceut.9b00182.

Prediction of Orthosteric and Allosteric Regulations on Cannabinoid Receptors Using Supervised Machine Learning Classifiers

Yuemin Bian^{†,‡,§}, Yankang Jing^{†,‡,§}, Lirong Wang^{†,‡,§}, Shifan Ma^{†,‡,§}, Jaden Jungho Jun^{†,‡,§}, Xiang-Qun Xie^{*,†,‡,§,||}

[†]Department of Pharmaceutical Sciences and Computational Chemical Genomics Screening Center, School of Pharmacy, School of Medicine, University of Pittsburgh, 335 Sutherland Drive, 206 Salk Pavilion, Pittsburgh, Pennsylvania 15261, United States

[‡]NIH National Center of Excellence for Computational Drug Abuse Research, School of Medicine, University of Pittsburgh, 335 Sutherland Drive, 206 Salk Pavilion, Pittsburgh, Pennsylvania 15261, United States

[§]Drug Discovery Institute, School of Medicine, University of Pittsburgh, 335 Sutherland Drive, 206 Salk Pavilion, Pittsburgh, Pennsylvania 15261, United States

^{||}Departments of Computational Biology and Structural Biology, School of Medicine, University of Pittsburgh, 335 Sutherland Drive, 206 Salk Pavilion, Pittsburgh, Pennsylvania 15261, United States

Abstract

Designing highly selective compounds to protein subtypes and developing allosteric modulators targeting them are critical considerations to both drug discovery and mechanism studies for cannabinoid receptors. It is challenging but in demand to have classifiers to identify active ligands from inactive or random compounds and distinguish allosteric modulators from orthosteric ligands. In this study, supervised machine learning classifiers were built for two subtypes of cannabinoid receptors, CB1 and CB2. Three types of features, including molecular descriptors, MACCS fingerprints, and ECFP6 fingerprints, were calculated to evaluate the compound sets from diverse aspects. Deep neural networks, as well as conventional machine learning algorithms including support vector machine, naïve Bayes, logistic regression, and ensemble learning, were applied. Their performances on the classification with different types of features were compared and discussed. According to the receiver operating characteristic curves and the calculated metrics, the advantages and drawbacks of each algorithm were investigated. The feature ranking was

*Corresponding Author: xix15@pitt.edu. Tel: 412-383-5276. Fax: 412-383-7436.

Supporting Information

The Supporting Information is available free of charge on the [ACS Publications website](https://pubs.acs.org) at DOI: [10.1021/acs.molpharmaceut.9b00182](https://doi.org/10.1021/acs.molpharmaceut.9b00182). Ranked molecular descriptor-based prediction scores for each machine learning algorithm by metrics (Tables S1–S3); ranked MACCS fingerprint-based prediction scores (Tables S4–S6); ranked ECFP6 fingerprint-based prediction scores (Tables S7–S9); and ROC curves for SVM models (Figure S1), ROC curves for MLP models (Figure S2), ROC curves for molecular descriptor-based prediction models (Figures S3–S5), ROC curves for MACCS-based prediction models (Figures S6–S8), and ROC curves for ECFP6-based prediction models (Figures S9–S11) (PDF)

The authors declare no competing financial interest.

followed to help extract useful knowledge about critical molecular properties, substructural keys, and circular fingerprints. The extracted features will then facilitate the research on cannabinoid receptors by providing guidance on preferred properties for compound modification and novel scaffold design. Besides using conventional molecular docking studies for compound virtual screening, machine-learning-based decision-making models provide alternative options. This study can be of value to the application of machine learning in the area of drug discovery and compound development.

Graphical Abstract

Keywords

cannabinoid receptor; allosteric regulation; machine learning; deep neural network; drug design

INTRODUCTION

Cannabis has been used for medical and recreational purposes more than 4000 years,^{1,2} and its medical use draws increasing attention nowadays.³ Remarkably, in June 2018, the US Food and Drug Administration (FDA) approved the cannabidiol to treat Lennox–Gastaut syndrome and Dravet syndrome, two rare and severe forms of epilepsy. It is the first FDA-approved drug that comprised an active ingredient derived from marijuana. There are generally two subtypes of cannabinoid receptors, termed CB1 and CB2, which share about 48% similarities on protein sequences⁴ but have distinctive distributions around the human body.^{5,6} CB1 is mainly expressed in the central nervous system⁶ associated with anxiety responses,⁷ drug addictions,⁸ motor controls,⁹ cardiovascular activities,¹⁰ and olfaction,¹¹ whereas CB2 is mainly expressed in the peripheral parts including the immune system and hematopoietic cells.¹² Therefore, targeting CB2 shows treatment benefits in auto-immune disorders, chronic inflammatory pain, breast cancer, osteoporosis, and liver and gastrointestinal diseases.⁴ Even if there are discussions on the existence of additional cannabinoid receptors like GPR18,¹³ GPR55,¹⁴ and GPR119,^{15,16} it still requires huge efforts for the research on CB1 and CB2 receptors. For G-protein-coupled receptors (GPCRs), there are generally multiple allosteric binding pockets besides the traditional

orthosteric sites.¹⁷ The allosteric modulators may not directly trigger the physiological responses but can have the saturable influence on the orthosteric regulation.^{18,19} The allosteric modulators can have preferable safety profiles owing to the saturable ceiling effects.^{20,21} Meanwhile, modulators can achieve certain degrees of selectivity, as allosteric binding pockets were under less evolutionary pressure for conservation.^{20,21} Therefore, (1) designing highly selective CB1/CB2 ligands and (2) developing allosteric modulators toward each target are two critical subjects to both the novel drug discovery and mechanism studies. Conventional computational chemistry methodologies including homology modeling, molecular docking, and molecular dynamics simulation have been applied, as well as medicinal chemistry approaches, to address the abovementioned topics.^{17,22–29} However, challenges do exist especially on the accuracy of *in silico* screening, and there is a demand for tools to identify active ligands from inactive or random compounds and, furthermore, to distinguish allosteric modulators from orthosteric ligands.

The approach we adapted to address these subjects is machine learning. Machine learning is the study of methods to automatically detect patterns in data and then use these patterns to predict future data or facilitate decision-making under uncertainty.³⁰ There are two main advantages of machine learning to benefit these topics. First, machine learning is capable of dealing with large data, showing a promising and active solution for the increased availability among cheminformatics data.³¹ Second, machine learning contains diverse algorithms to develop accurate predictive models and have been successfully used in many research areas as the driving force of artificial intelligence.³² Developing machine-learning-based virtual screening pipelines to mine the large databases for potential hits toward the target proteins brings opportunities to the field of drug discovery. The substructural analysis proposed by Cramer et al.³³ as a method for the automated analysis of biological data in 1974 was considered as the first application of machine learning in drug discovery. In recent years, Li et al. reported a multitask deep autoencoder neural network model to predict the human cytochrome P450 inhibition.³⁴ Korotcov et al. constructed a series of machine learning models on diverse drug discovery datasets to conduct a systematic comparison of the model's performances.³⁵ Significantly, the AlphaFold from the DeepMind recently won the CASP13, a biennial assessment of protein structure prediction methods, using deep learning approaches.

Our group previously reported machine learning models for ligand selectivity and biological activity predictions.^{36,37} In the current study, we extended our scope by including diverse types of descriptors and multiple machine learning algorithms for model training. With the focus on identifying active cannabinoid ligands from inactive or random compounds and distinguishing CB1 allosteric modulators from orthosteric ligands, three specific compound sets were created through data intergradation. Three types of features were calculated to evaluate the compound sets through various aspects. Seven machine learning algorithms were applied to generate classifier models. Metrics were referred to evaluate the model performances. The feature ranking was followed to help identify critical features that may provide guidance on compound modification and novel compound design on cannabinoid receptors afterward. The authors explored the combinations of different types of molecular features and machine learning algorithms, which can result in a robust virtual screening method for research regarding cannabinoid receptors. To the best of our knowledge, this

study gave the first report on the successful application on classifying GPCR orthosteric and allosteric ligands using machine learning algorithms. The study can also demonstrate the value of building and applying machine-learning-based decision-making models to benefit the studies on cheminformatics and drug discovery.

EXPERIMENTAL SECTION

Dataset Preparation.

Chemical information from diverse drug discovery databases was combined to generate the CB1 active compound–inactive/random compound (CB1) set, CB2 active compound–inactive/random compound (CB2) set, and CB1 orthosteric compound–allosteric compound (CB1O/CB1A) set. The ChEMBL database³⁸ was referred for collecting orthosteric ligands with experimental K_i values for both CB1 and CB2 receptors. The ZINC database³⁹ was referred for collecting druglike random compounds to function as decoys to give “white noise”. The allosteric database (ASD)⁴⁰ was referred for collecting CB1 allosteric modulators. The cutoff of K_i value to distinguish active and inactive compounds was set to be 100 nM. The cutoff for the mutual similarity of compounds within a dataset was set to be 0.8. The similarity was measured by the Tanimoto coefficient through MACCS fingerprints. Five thousand clean druglike compounds were integrated to both the CB1 and CB2 datasets to mix with inactive compounds. The compound set for CB2 orthosteric compound–allosteric compound was not generated mainly due to the limited amount of CB2 allosteric modulators available. Once only limited input data are available, the machine will be incapable to detect patterns in the data and then these patterns cannot be used to make future predictions. The developed datasets underwent stratified splitting to result in a training (80%) set and test (20%) set, while the ratios of active and inactive/random (orthosteric and allosteric) compounds maintain equal proportions in each split set. The KNIME software⁴¹ was applied in input data intergradation, fingerprint-based similarity calculation, and labeling.

Descriptor Calculation.

Both physical–chemical descriptors and molecular fingerprints were used to represent the molecular structure of all compounds in the three compound sets. For physical–chemical molecular descriptors, 119 molecular descriptors, including ExactMW, SlogP, TPSA, NumHBD, NumHBA, etc., were calculated using RDKit.⁴² For molecular fingerprints, MACCS fingerprints and ECFP6 fingerprints were calculated with a CDK toolkit.⁴³ MACCS fingerprints have 166 binary fingerprints as substructure keys, each of which indicates the presence of one of the 166 MDL MACCS structural keys calculated from the molecular graph. ECFP6 are circular topological fingerprints with 1024 descriptors. The ECFP6 represent molecular structures by means of circular atom neighborhoods. The features represent the presence of particular substructures.

Machine Learning.

A prediction pipeline was developed for supervised classification with various machine learning algorithms, including support vector machine (SVM),⁴⁴ neural network/multilayer perceptron (MLP),⁴⁵ random forest (RF),⁴⁶ AdaBoost decision tree (ABDT),⁴⁷ decision tree

(DT),⁴⁸ naïve Bayes (NB),⁴⁹ and logistic regression.⁵⁰ Open-source Python module Scikit-learn⁵¹ was used for model training, data prediction, and result interpretation.

Support vector machine (SVM) is effective in high-dimensional spaces. In cases where the number of samples is less than the number of dimensions, SVM can be effective in using different kernel functions for the decision function to handle the problems. The *svm.SVC()* method with three kernel functions (*linear*, *rbf*, *poly*) from Scikit-learn was applied. The parameter *probability* was set to true. The parameter *random_state* was set to *random_state*. The SVM model with the best performance was saved after the optimization on penalty parameter *C* and parameter γ for *rbf* and *poly* kernels.

Multilayer perceptron (MLP) is a supervised learning algorithm that has the capacity to learn nonlinear models in real time. MLP can have one or more nonlinear hidden layers between the input and output layers. For each hidden layer, different numbers of hidden neurons can be assigned. Each hidden neuron gives a weighted linear summation for the values from the previous layer, and the nonlinear activation function is followed. The output values are reported after the output layer transforms the values from the last hidden layer. The *MLPClassifier()* method in Scikit-learn with one to five hidden layers and a constant learning rate was applied. The number of hidden neurons for each hidden layer was set to be constant to the number of input features. The solver for the weight optimization was set to *adam* for CB1 and CB2 datasets in the observation of the relatively large datasets (thousands of samples) involved and *lbfgs* for the CB10/CB1A dataset. The following parameters were optimized before the model training: activation function (*identity*, *logistic*, *tanh*, *relu*), L2 penalty alpha (1×10^{-2} , 1×10^{-3} , 1×10^{-4} , 1×10^{-5}), and learning rate (0.1, 0.01, 0.001, 0.0001).

Random forest (RF) is an ensemble method that combines the predictions of a number of decision tree classifiers to improve the robustness over a single estimator. As the averaging method, the driving principle of RF is to average predictions after independently building several estimators. *RandomForestClassifier()* was applied with parameter *bootstrap* set to true. The model was saved after the optimization on parameters *n_estimators* (10, 100, 1000) and *max_depth* (2, 3, 4, 5).

AdaBoost decision tree (ABDT) is another ensemble method. Different from the averaging methods, the boosting methods have the estimators built sequentially and each one tries to reduce the bias of the combined estimator. The decision tree models are combined in ABDT to produce a powerful ensemble. *AdaBoostClassifier()* was applied with the optimization on parameters *n_estimators* (10, 100, 1000) and *learning_rate* (0.01, 0.1, 1).

Decision tree (DT) is a nonparametric supervised learning method to build models that can learn decision rules from the input data and make predictions on the values of a target variable. DT can have trees visualized, which is simple to understand and interpret. *DecisionTreeClassifier()* was applied for generating models with the optimization on parameter *max_depth*.

Naïve Bayes (NB) algorithms are supervised learning methods that have an assumption of conditional independence between every pair of features. NB algorithms are based on

applying Bayes' theorem. *GaussianNB()*, which implements the Gaussian naïve Bayes algorithm for classification, was applied for datasets with molecular descriptors as features. The likelihood of the features is assumed to be Gaussian. *BernoulliNB()*, which implements the training and classification algorithms for data that follows multivariate Bernoulli distributions, was applied for datasets with fingerprints as features. Given that Bernoulli naïve Bayes requires binary-valued feature vectors for samples. The prior probabilities of the classes were set to none.

Logistic regression is a linear model for classification rather than regression. The logistic function was used to model the probabilities, which describe the possible outcomes of one single trial. *LogisticRegression()* was applied to implement the algorithm with l2 penalty. The parameter solver was set to sag to handle the multinomial loss in large datasets.

Model Evaluation.

Sixfold cross-validation for each of nine combinations of datasets and descriptor types was performed for model generation and evaluation. The Scikit-learn module StratifiedKFold was used to split the dataset into 6-folds. The model was trained using 5-folds as training data, and the resulting model is validated on the remaining fold of data. A series of metrics were calculated to evaluate the performance of machine learning models from diverse aspects. Model evaluation and feature selection functions in the Python module Scikit-learn were applied for the calculation. Python module matplotlib⁵² was used in plotting.

Area under the receiver operating characteristic (ROC) curve (AUC) was calculated with *auc()* after true-positive rate and false-positive rate have been acquired with *roc_curve()*. AUC computes the area under the receiver operating characteristic (ROC) curve using the trapezoidal rule. AUC can be referred to indicate the performance of the model on separating classes.

Balanced *F*-score or *F*-measure (F_1 score) was calculated with *f1_score()*. The F_1 score can be interpreted as the weighted average of the precision and the recall. The precision and the recall have relatively equal contribution to the F_1 score. $F_1 = 2 \times (\text{precision} \times \text{recall}) / (\text{precision} + \text{recall})$.

The accuracy classification score (ACC) was calculated with *accuracy_score()*. ACC computes subset accuracy that whether the label predicted for one sample matches with the corresponding true value.

Cohen's κ was calculated with *cohen_kappa_score()*. Cohen's κ measures interannotator agreement, which expresses the level of agreement between two annotators on a classification problem.

Matthew's correlation coefficient (MCC) was calculated with *Matthews_corrcoef()*. MCC is used to measure the quality of binary and multiclass classifications. It is a balanced measure that both the true and false positives and negatives are considered.

Precision was calculated with *precision_score()*. The precision measures the ability of a model to not label a negative sample as positive. Precision = true positives/(true positives + false positives).

Recall was calculated with *recall_score()*. The recall measures the ability of a model to find out all of the positive samples. Recall = true positives/(true positive + false negative).

Feature Ranking.

Recursive feature elimination (RFE) from *sklearn.feature_selection* was implemented for feature ranking. The *n_features_to_select* was set to 1. The *step* was set to 1. The RFE is an iterative process to consider a smaller set of features. The weights were assigned to features. The importance of features is analyzed, and the least important features are pruned. RDKit molecular descriptors (119) were plotted into a 7×17 matrix. The least important feature from the 166 MACCS fingerprint features was first dropped, and the remaining 165 features were plotted into an 11×15 matrix. ECFP6 fingerprint features (1024) were plotted into a 32×32 matrix. Python module matplotlib was used in plotting.

RESULTS AND DISCUSSION

Overall Workflow.

The schematic illustration on the workflow of this study is shown in Figure 1. CB1 and CB2 compounds with experimental K_i values were extracted from the ChEMBL database. The activity cutoff was set to 100 nM to distinguish active and inactive compounds. Druglike compounds were randomly selected from the ZINC database to represent a larger chemical space. CB1 allosteric modulators were then collected from the ASD. The duplicated and similar (Tanimoto coefficient over 0.8 based on MACCS fingerprints) compounds were first filtered out for CB1 active, CB1 inactive, CB2 active, CB2 inactive, CB1 allosteric, and random compounds. The random compounds were then mixed with CB1 inactive and CB2 inactive compounds. Three compound sets, CB1 active/CB1 inactive and random compounds (CB1), CB2 active/CB2 inactive and random compounds (CB2), and CB1 orthosteric/CB1 allosteric compounds (CB1O/CB1A), were created by integrating the compounds described above (Table 1). Three types of features, molecular descriptors, MACCS fingerprints (structural keys), and ECFP6 fingerprints (circular), were calculated for the three compound sets to result in nine datasets: (1) CB1 descriptors, (2) CB1 MACCS, (3) CB1 ECFP6, (4) CB2 descriptors, (5) CB2 MACCS, (6) CB2 ECFP6, (7) CB1O/CB1A descriptors, (8) CB1O/CB1A MACCS, and (9) CB1O/CB1A ECFP6. Active compounds (or CB1 orthosteric compounds) and inactive and random compounds (or CB1 allosteric compounds) were labeled for classification. The training sets and test sets were divided at an 80:20 ratio for all of these nine datasets. Seven supervised machine learning algorithms were applied to build classifiers for each of the prepared dataset, which resulted in 99 classifier models to identify (1) active from inactive and random compounds and (2) CB1 orthosteric from CB1 allosteric compounds. Different types of features can evaluate the properties of compounds from diverse aspects, and different machine learning algorithms may favor distinctive data structures. In this case, three compounds sets were calculated by 3 types of

features and evaluated with 11 algorithms to better cover the possible combinations and perform the classification. The detailed processes are specified in the Experimental Section.

Prediction Results.

The AUC values of all machine learning models (seven algorithms on nine datasets) for the training set and test set are summarized in Tables 2 and 3. Models gave consistent performances on both the training set and test set. The NB outperformed all of the other algorithms for three times on datasets CB1 descriptors (Gaussian NB), CB1 ECFP6 (Bernoulli NB), and CB2 ECFP6 (Bernoulli NB). The logistic regression received the largest AUC for two times on datasets CB1 MACCS and CB2 MACCS. The MLP with multiple hidden layers achieved the best performances on small fingerprint-based datasets, CB1O/CB1A MACCS (ABDT achieved the highest AUC value on the test set) and CB1O/CB1A ECFP6. The SVM and ABDT each scored the highest once on datasets CB2 descriptor and CB1O/CB1A descriptors, respectively.

Figure 2 shows the ROC curves for these nine best-performing models on each dataset. Considering that (1) the size of the compound sets varies (relatively large datasets for CB1 and CB2 with about 6000 entries each, and small dataset for CB1O/CB1A with about 600 entries) and that (2) 3 different types of features for the compound sets were calculated through diverse approaches, the constructed 9 datasets can have distinctive data structures. The performance of a certain machine learning algorithm can be influenced by the structure of the input data. This can explain that certain algorithms can outperform others with some datasets while becoming inferior predictors in the others.

Gaussian NB can be very fast on the classification and is suitable for the discrete data in dataset CB1 descriptors (Figure 2A). Bernoulli NB demonstrates its ability to building classifiers to handle large binary datasets among datasets CB1 ECFP6 and CB2 ECFP6 (Figure 2G,H). Three kernel functions (*linear*, *poly*, *rbf*) were adopted for SVM in this study (Supporting Information, Figure S1). The linear SVM classifier gave the best performance on dataset CB2 descriptors (Figure 2B). The ensemble method ABDT that has the DTs built in sequence to reduce the bias and increase the prediction power scored highest on dataset CB1O/CB1A descriptors (Figure 2C). There is no surprise that both the ensemble models with the averaging method (RF) and boosting method (ABDT) improved the weak classifier (DT) across all of these nine datasets (Table 2). Similar to Bernoulli NB, the logistic regression can have superior performance on large binary datasets and was the best in class for datasets CB1 MACCS and CB2 MACCS (Figure 2D,E). The performance can be partially contributed by the sag as the solver, which handles the multinomial loss in large datasets. When it comes to relatively small binary datasets, CB1O/CB1A MACCS and CB1O/CB1A ECFP6, the MLP with multiple hidden layers demonstrated its advantages. Usually, neural networks (NNs) with three or more hidden layers are considered as deep neural networks (DNNs). The DNN can be efficient in building classifiers to summarize high-dimensional information with relatively fewer input samples. While the DNN can outperform the shallow NN in certain cases, the determination of the number of hidden layers and the number of hidden neurons for each layer can be an iterative process since

there is no given reference (Supporting Information, Figure S2). The ROC curves for all of the models are attached as Supporting Information figures.

Model Evaluation with Metrics.

Instead of using only the AUC score, a series of metrics were calculated to further explore the performance of each machine learning algorithm on different feature types. Metrics functions assess prediction errors for specific purposes and can evaluate the model performance from various aspects. The other metrics involved in this study are F_1 score, ACC, Cohen's κ , MCC, precision, and recall. The mean score was then calculated by averaging all of the individual metrics.

The metrics scores for the classifier models with molecular descriptors as features were averaged among datasets CB1 descriptors, CB2 descriptors, and CB1O/CB1A descriptors (Table 4). The ABDT model outperformed the others with the highest scores on AUC, F_1 score, ACC, Cohen's κ , and MCC. The MLP model with three hidden layers was also favored with the top precision. The Gaussian NB had the best score on recall but moderate scores for the other metrics. The metrics scores for the classifier models with MACCS fingerprints as features were averaged among datasets CB1 MACCS, CB2 MACCS, and CB1O/CB1A MACCS (Table 5). The classic MLP model with one hidden layer ranked the top with the highest scores on AUC, Cohen's κ , MCC, and precision. The DNN models, especially the model with four hidden layers, can have comparable but inferior scores, which demonstrate that a better performance is not guaranteed when a model goes deeper. The model selection is more a case-by-case analysis based on the structure of the input data. The logistic regression ranked second to MLPs with the best scores on recall. The ensemble method ABDT achieved the highest scores on ACC. The metrics scores for the classifier models with ECFP6 fingerprints as features were averaged among datasets CB1 ECFP6, CB2 ECFP6, and CB1O/CB1A ECFP6 (Table 6). The logistic regression had the highest scores for ACC, Cohen's κ , MCC, and recall and ranked the top. The Bernoulli NB ranked second with the top scores on AUC and F_1 score. The MLPs achieved moderate scores, while the MLP model with four hidden layers received the highest score on precision. The metrics tables for all of the models are attached as Supporting Information tables.

One trend can be observed. Over the three tables, there were high scores for AUC, ACC, and recall and moderate scores for F_1 score, Cohen's κ , MCC, and precision, which deserve the attention. The F_1 score can be interpreted as the weighted average of the precision and recall, which can then be affected by a low precision. The low precision indicates the relatively high false-positive rate for the classification. Also, the MCC and Cohen's κ can be affected by this high false-positive rate given that MCC is a balanced measure that both the true and false positives and negatives are considered, and Cohen's κ measures interannotator agreement. The cause of the high false-positive rate was the mixed classification of random compounds and the inactive compounds with the same label of negative. The random compounds are supposed to have a low hit rate (~0%) of actives for both the cannabinoid targets, but the hits can still exist. Even though 80% of the random compounds were grouped as the training set, the characteristics can hardly be summarized to classify the remaining random compounds in the test set, given that they have random structures with random

scaffolds. The false-positive rate increases once the random compounds fulfill the rules for the actives and been classified as active by the algorithms. This imbalance on scores can be observed from all of the datasets with the integration of random compounds to inactives (Supporting Information, Tables S1, S2, S4, S5, S7, and S8). For the CB1O/CB1A datasets, only the CB1 orthosteric and allosteric ligands are collected. No imbalance can be observed among the metrics calculated in this case (Supporting Information, Tables S3, S6, and S9). The random druglike compounds extended the chemical spaces in a dataset dramatically. But the additional attention will need to be paid regarding the increased false-positive rate on supervised model prediction at the same time.

Feature Ranking.

The feature ranking was followed using the recursive feature elimination (Figure 3). The contribution of each feature on making a correct prediction was obtained through either the *coef* attribute or the *feature_importances* attribute. The importance of features on the classification can vary dramatically. Giving ranks to molecular descriptors or fingerprint features can help identify vital molecular properties and key substructures that are critical for the algorithms to make the decision on the classification. The identified molecular properties can then provide guidance on the direction of compound modifications. While the identified vital substructures can be referred to as potential materials to build novel scaffolds or substitutions based on known structures. Also, the critical substructures themselves and the combination of them may contribute to a target-specific fragment database, which may facilitate the structure-based and fragment-based drug design afterward.

The feature ranking for molecular descriptors was based on the ABDT models (Figure 3A–C) since the ABDT had the highest overall scores according to the metrics calculation. The 119 features were plotted into a 7×17 matrix. Distinctive matrix patterns can be observed among the three datasets, which indicates that features had different weights in these models and the CB1 and CB2 compounds share diverse molecular properties. The feature ranking for MACCS fingerprints was based on the logistic regression models (Figure 3D–F). The MLP model with one hidden layer was favored by the metrics calculation. However, the nature of hidden layers in neural networks has neither *coef* nor *feature_importances* attributes, which disabled the RFE analysis. With the least important feature been deleted, 165 MACCS features were plotted into an 11×15 matrix for visualization. Similar matrix patterns can be traced among the three datasets that the majority of features on the first row gave weak contributions to the final classification. Both active and inactive/random compounds can share these similar or identical substructures. It indicates that (1) these substructures are primary components in compound formation or (2) these substructures can have negligible effects on cannabinoid receptor binding. The feature ranking for ECFP6 fingerprints was also based on the logistic regression models (Figure 3G–I) as they ranked top on metrics calculation. ECFP6 features (1024) were plotted into a 32×32 matrix. The random patterns can be the result of the circular atom neighborhoods. However, again, random patterns also suggest that distinctive rules were adopted based on the diverse structural properties and were applied to classify active (orthosteric) and inactive/random (allosteric) compounds in each dataset.

To demonstrate further how this study can facilitate the orthosteric and allosteric molecule designs, we analyzed and listed the important molecular descriptors (Table 7) and MACCS fingerprints (Table 8) that ranked top 10 for the classification of orthosteric and allosteric ligands. As shown in Table 7, the distributions of each feature can vary between orthosteric and allosteric ligands. For example, the differences between the skewness and kurtosis of feature `smr_VSA7` indicate two distinctive distributions for CB1O and CB1A. It is foreseeable that the difference may not be significant for one single feature (otherwise, one feature itself is good enough to distinguish allosteric modulators from orthosteric ligands), but each feature gives its contribution to the classification. Table 8 lists the top 10 substructure keys in the logistic regression model. Features 144, 76, 130, 84, 9, 81, and 13 are favored by CB1 allosteric ligands. For example, feature 144 can be interpreted as the substructure of amide (with an aromatic query bond) from the example compound with CAS ID 1207203-33-5. Of 208 allosteric ligands, 19 have this feature, while only 3 out of 376 orthosteric ligands have this feature. Features 115, 78, and 145 are favored by CB1 orthosteric ligands. For example, feature 115 can be interpreted as the substructure of a methyl group connected with a methylene group through any valid periodic table elements. Of 376 orthosteric ligands, 80 have this feature, while only 27 out of 208 allosteric ligands have this feature. The full list of MACCS fingerprint keys is detailed in the Supporting Information. These features contribute to the compound classification and can be associated with the specific receptor–ligand interactions and the target selectivity.

CONCLUSIONS

In this study, supervised machine learning classifiers were built to predict the orthosteric ligands and allosteric modulators for cannabinoid receptors. Three types of features were calculated to evaluate the compound sets from diverse aspects, including molecular descriptors and fingerprints. Seven machine learning algorithms were applied to build classifier models. The performances of algorithms on different types of features were compared and discussed. With the ROC curves and the calculated metrics, the advantages and drawbacks for each specific algorithm were investigated. The feature ranking was followed to help identify critical molecular properties, key substructures, and circular fingerprints that may provide guidance on compound modification and novel structure design for cannabinoid receptors afterward. To the best of our knowledge, this study is the first to report the successful application on classifying GPCR orthosteric and allosteric ligands using machine learning algorithms. In a nutshell, the developed machine-learning-based decision-making models provide additional choices on compound screening besides the conventional *in silico* methods like molecular docking studies and molecular pharmacophore models. The benefit of this study may not only be limited to the research regarding cannabinoid receptors but also be of value to the application of machine learning in the area of drug discovery and compound development.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

The authors would like to acknowledge the funding support to the Xie laboratory from the NIH NIDA (P30 DA035778A1) and DOD (W81XWH-16-1-0490).

REFERENCES

- (1). Hall W; Degenhardt L Adverse health effects of non-medical cannabis use. *Lancet* 2009, 374, 1383–1391. [PubMed: 19837255]
- (2). Bian Y.-m.; He X.-b.; Jing Y.-k.; Wang L.-r.; Wang J.-m.; Xie X-Q Computational systems pharmacology analysis of cannabidiol: a combination of chemogenomics-knowledgebase network analysis and integrated in silico modeling and simulation. *Acta Pharmacol. Sin* 2019, 40, 374–386. [PubMed: 30202014]
- (3). Hill KP Medical marijuana for treatment of chronic pain and other medical and psychiatric problems: a clinical review. *J. Am. Med. Assoc* 2015, 313, 2474–2483.
- (4). Yang P; Wang L; Xie X-Q Latest advances in novel cannabinoid CB2 ligands for drug abuse and their therapeutic potential. *Future Med. Chem* 2012, 4, 187–204. [PubMed: 22300098]
- (5). Aghazadeh Tabrizi M; Baraldi PG; Borea PA; Varani K Medicinal chemistry, pharmacology, and potential therapeutic benefits of cannabinoid CB2 receptor agonists. *Chem. Rev* 2016, 116, 519–560. [PubMed: 26741146]
- (6). Mackie K Distribution of Cannabinoid Receptors in the Central and Peripheral Nervous System In *Cannabinoids*; Springer, 2005; pp 299–325.
- (7). Hill MN; McLaughlin RJ; Bingham B; Shrestha L; Lee TT; Gray JM; Hillard CJ; Gorzalka BB; Viau V Endogenous cannabinoid signaling is essential for stress adaptation. *Proc. Natl. Acad. Sci. U.S.A* 2010, 107, 9406–9411. [PubMed: 20439721]
- (8). De Vries TJ; Schoffelmeer AN Cannabinoid CB1 receptors control conditioned drug seeking. *Trends Pharmacol. Sci* 2005, 26, 420–426. [PubMed: 15992935]
- (9). McLaughlin PJ; Delevan CE; Carnicom S; Robinson JK; Brener J Fine motor control in rats is disrupted by delta-9-tetrahydrocannabinol. *Pharmacol., Biochem. Behav* 2000, 66, 803–809. [PubMed: 10973519]
- (10). Varga K; Wagner JA; Bridgen DT; Kunos G Platelet-and macrophage-derived endogenous cannabinoids are involved in endotoxin-induced hypotension. *FASEB J* 1998, 12, 1035–1044. [PubMed: 9707176]
- (11). Elphick MR; Egertova M The neurobiology and evolution of cannabinoid signalling. *Philos. Trans. R. Soc., B* 2001, 356, No. 381.
- (12). Schatz AR; Lee M; Condie RB; Pulaski JT; Kaminski NE Cannabinoid receptors CB1 and CB2: a characterization of expression and adenylate cyclase modulation within the immune system. *Toxicol. Appl. Pharmacol* 1997, 142, 278–287. [PubMed: 9070350]
- (13). Kohno M; Hasegawa H; Inoue A; Muraoka M; Miyazaki T; Oka K; Yasukawa M Identification of N-arachidonylglycine as the endogenous ligand for orphan G-protein-coupled receptor GPR18. *Biochem. Biophys. Res. Commun* 2006, 347, 827–832. [PubMed: 16844083]
- (14). Ryberg E; Larsson N; Sjögren S; Hjorth S; Hermansson NO; Leonova J; Elebring T; Nilsson K; Drmota T; Greasley P The orphan receptor GPR55 is a novel cannabinoid receptor. *Br. J. Pharmacol* 2007, 152, 1092–1101. [PubMed: 17876302]
- (15). Overton HA; Fyfe M; Reynet C GPR119, a novel G protein-coupled receptor target for the treatment of type 2 diabetes and obesity. *Br. J. Pharmacol* 2008, 153, S76–S81. [PubMed: 18037923]
- (16). Brown AJ Novel cannabinoid receptors. *Br. J. Pharmacol* 2007, 152, 567–575. [PubMed: 17906678]
- (17). Bian Y; Feng Z; Yang P; Xie X-Q Integrated in silico fragment-based drug design: case study with allosteric modulators on metabotropic glutamate receptor 5. *AAPS J* 2017, 19, 1235–1248. [PubMed: 28560482]

- (18). Morales P; Goya P; Jagerovic N; Hernandez-Folgado L Allosteric modulators of the CB1 cannabinoid receptor: a structural update review. *Cannabis Cannabinoid Res* 2016, 1, 22–30. [PubMed: 28861476]
- (19). Khurana L; Mackie K; Piomelli D; Kendall DA Modulation of CB1 cannabinoid receptor by allosteric ligands: pharmacology and therapeutic opportunities. *Neuropharmacology* 2017, 124, 3–12. [PubMed: 28527758]
- (20). Conn PJ; Christopoulos A; Lindsley CW Allosteric modulators of GPCRs: a novel approach for the treatment of CNS disorders. *Nat. Rev. Drug Discovery* 2009, 8, 41–54. [PubMed: 19116626]
- (21). Wenthur CJ; Gentry PR; Mathews TP; Lindsley CW Drugs for allosteric sites on receptors. *Annu. Rev. Pharmacol. Toxicol* 2014, 54, 165–184. [PubMed: 24111540]
- (22). Yang P; Wang L; Feng R; Almhazia AA; Tong Q; Myint K-Z; Ouyang Q; Alqarni MH; Wang L; Xie X-Q Novel triaryl sulfonamide derivatives as selective cannabinoid receptor 2 inverse agonists and osteoclast inhibitors: discovery, optimization, and biological evaluation. *J. Med. Chem* 2013, 56, 2045–2058. [PubMed: 23406429]
- (23). Yang P; Myint K-Z; Tong Q; Feng R; Cao H; Almhazia AA; Alqarni MH; Wang L; Bartlow P; Gao Y; et al. Lead discovery, chemistry optimization, and biological evaluation studies of novel biamide derivatives as CB2 receptor inverse agonists and osteoclast inhibitors. *J. Med. Chem* 2012, 55, 9973–9987. [PubMed: 23072339]
- (24). Wang H; Duffy RA; Boykow GC; Chackalamannil S; Madison VS Identification of novel cannabinoid CB1 receptor antagonists by using virtual screening with a pharmacophore model. *J. Med. Chem* 2008, 51, 2439–2446. [PubMed: 18363352]
- (25). Evers A; Klabunde T Structure-based drug discovery using GPCR homology modeling: successful virtual screening for antagonists of the alpha1A adrenergic receptor. *J. Med. Chem* 2005, 48, 1088–1097. [PubMed: 15715476]
- (26). Gianella-Borradori M; Christou I; Bataille CJ; Cross RL; Wynne GM; Greaves DR; Russell AJ Ligand-based virtual screening identifies a family of selective cannabinoid receptor 2 agonists. *Bioorg. Med. Chem* 2015, 23, 241–263. [PubMed: 25487422]
- (27). Bian Y; Xie X-QS Computational Fragment-Based Drug Design: Current Trends, Strategies, and Applications. *AAPS J* 2018, 20, No. 59.
- (28). Gado F; Di Cesare Mannelli L; Lucarini E; Bertini S; Cappelli E; Digiaco M; Stevenson LA; Macchia M; Tuccinardi T; Ghelardini C; et al. Identification of the first synthetic allosteric modulator of the CB2 receptors and evidence of its efficacy for neuropathic pain relief. *J. Med. Chem* 2019, 62, 276–287. [PubMed: 29990428]
- (29). Petrucci V; Chicca A; Glasmacher S; Paloczi J; Cao Z; Pacher P; Gertsch J Pepcan-12 (RVD-hemopressin) is a CB2 receptor positive allosteric modulator constitutively secreted by adrenals and in liver upon tissue damage. *Sci. Rep* 2017, 7, No. 9560.
- (30). Robert C Redefining the Performance Auditing Space; Taylor & Francis, 2014.
- (31). Jing Y; Bian Y; Hu Z; Wang L; Xie X-QS Deep learning for drug design: an artificial intelligence paradigm for drug discovery in the big data era. *AAPS J* 2018, 20, No. 58.
- (32). Kotsiantis SB; Zaharakis I; Pintelas P Supervised Machine Learning: A Review of Classification Techniques In Emerging Artificial Intelligence Applications in Computer Engineering; IOS Press, 2007; Vol. 160, pp 3–24.
- (33). Cramer RD III; Redl G; Berkoff CE Substructural analysis. Novel approach to the problem of drug design. *J. Med. Chem* 1974, 17, 533–535. [PubMed: 4830256]
- (34). Li X; Xu Y; Lai L; Pei J Prediction of human cytochrome P450 inhibition using a multi-task deep autoencoder neural network. *Mol. Pharmaceutics* 2018, 15, 4336–4345.
- (35). Korotcov A; Tkachenko V; Russo DP; Ekins S Comparison of deep learning with multiple machine learning methods and metrics using diverse drug discovery data sets. *Mol. Pharmaceutics* 2017, 14, 4462–4475.
- (36). Myint K-Z; Wang L; Tong Q; Xie X-Q Molecular fingerprint-based artificial neural networks QSAR for ligand biological activity predictions. *Mol. Pharmaceutics* 2012, 9, 2912–2923.
- (37). Ma C; Wang L; Yang P; Myint KZ; Xie X-Q LiCABEDSII. Modeling of ligand selectivity for G-protein-coupled cannabinoid receptors. *J. Chem. Inf. Model* 2013, 53, 11–26. [PubMed: 23278450]

- (38). Gaulton A; Hersey A; Nowotka M; Bento AP; Chambers J; Mendez D; Mutowo P; Atkinson F; Bellis LJ; Cibrián-Uhalte E; et al. The ChEMBL database in 2017. *Nucleic Acids Res* 2017, 45, D945–D954. [PubMed: 27899562]
- (39). Irwin JJ; Shoichet BK ZINC—A free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model* 2005, 45, 177–182. [PubMed: 15667143]
- (40). Shen Q; Wang G; Li S; Liu X; Lu S; Chen Z; Song K; Yan J; Geng L; Huang Z; et al. ASD v3. 0: unraveling allosteric regulation with structural mechanisms and biological networks. *Nucleic Acids Res* 2016, 44, D527–D535. [PubMed: 26365237]
- (41). Berthold MR; Cebron N; Dill F; Gabriel TR; Kötter T; Meinl T; Ohl P; Thiel K; Wiswedel B KNIME—the Konstanz information miner: version 2.0 and beyond. *ACM SIGKDD Explor. Newsl* 2009, 11, 26–31.
- (42). Landrum G, RDKit: Open-source cheminformatics 2006, <http://www.rdkit.org>.
- (43). Steinbeck C; Hoppe C; Kuhn S; Floris M; Guha R; Willighagen EL Recent developments of the chemistry development kit (CDK)—an open-source java library for chemo- and bioinformatics. *Curr. Pharm. Des* 2006, 12, 2111–2120. [PubMed: 16796559]
- (44). Bennett KP; Campbell C Support vector machines: hype or hallelujah? *ACM SIGKDD Explor. Newsl* 2000, 2, 1–13.
- (45). Schmidhuber J Deep learning in neural networks: An overview. *Neural Networks* 2015, 61, 85–117. [PubMed: 25462637]
- (46). Breiman L Random forests. *Mach. Learn* 2001, 45, 5–32.
- (47). Freund Y; Schapire RE A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci* 1997, 55, 119–139.
- (48). Safavian SR; Landgrebe D A survey of decision tree classifier methodology. *IEEE Trans. Syst. Man Cybern* 1991, 21, 660–674.
- (49). John GH; Langley P Estimating Continuous Distributions in Bayesian Classifiers, *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*; Morgan Kaufmann Publishers Inc., 1995; pp 338–345.
- (50). Hosmer DW Jr.; Lemeshow S; Sturdivant RX *Applied Logistic Regression*; John Wiley & Sons, 2013; Vol. 398.
- (51). Pedregosa F; Varoquaux G; Gramfort A; Michel V; Thirion B; Grisel O; Blondel M; Prettenhofer P; Weiss R; Dubourg V Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res* 2011, 12, 2825–2830.
- (52). Hunter JD Matplotlib: A 2D graphics environment. *Comput. Sci. Eng* 2007, 9, 90–95.

Figure 1.
Overall workflow for data processing.

Figure 2.
ROC curves for best-performing models of each dataset.

Figure 3.
Feature ranking on three datasets.

Table 1.

Dataset Information in Detail

dataset	dataset references	total number	cutoff for active ligands (nM)	cutoff for mutual similarity	number of active ligands	number of inactive ligands	number of random compounds
CB1	using ChEMBL for active and inactive ligands; ZINC for random druglike compounds	5874	100	0.8	376	498	5000
CB2	using ChEMBL for active and inactive ligands; ZINC for random druglike compounds	5949	100	0.8	385	564	5000
CB1O/CB1A	using ChEMBL for orthosteric ligands; ASD for allosteric ligands	584		0.8	376 (orthosteric)	208 (allosteric)	

Table 2.

AUC Values of All Machine Learning Models with Each Dataset on Training Set^a

	datasets	MLP										logistic
		SVM	MLP_1	MLP_2	MLP_3	MLP_4	MLP_5	RF	ABDT	DT	NB	
molecular descriptor	CB1 train	0.926	0.886	0.870	0.875	0.903	0.916	0.931	0.924	0.879	0.940	0.866
	CB2 train	0.944	0.922	0.919	0.919	0.927	0.925	0.910	0.943	0.842	0.932	0.901
MACCS	CB10/CB1A train	0.940	0.918	0.757	0.849	0.908	0.913	0.914	0.982	0.819	0.886	0.908
	CB1 train	0.857	0.894	0.884	0.877	0.879	0.875	0.871	0.879	0.802	0.851	0.905
	CB2 train	0.924	0.935	0.919	0.925	0.927	0.923	0.896	0.919	0.828	0.884	0.938
ECFP6	CB10/CB1A train	0.935	0.953	0.958	0.962	0.963	0.953	0.889	0.961	0.818	0.870	0.939
	CB1 train	0.867	0.878	0.878	0.866	0.858	0.875	0.908	0.895	0.827	0.930	0.907
	CB2 train	0.923	0.922	0.928	0.931	0.925	0.921	0.909	0.925	0.821	0.945	0.932
	CB10/CB1A train	0.957	0.981	0.979	0.984	0.982	0.966	0.919	0.967	0.866	0.973	0.972

^aEach bold entry shows the highest metric value among the machine learning models using different algorithms.

Table 3.

AUC Values of All Machine Learning Models with Each Dataset on Test Set^a

datasets	MLP										logistic
	SVM	MLP_1	MLP_2	MLP_3	MLP_4	MLP_5	RF	ABDT	DT	NB	
molecular descriptor	0.922	0.852	0.854	0.866	0.898	0.891	0.914	0.915	0.818	0.935	0.826
CB1 test											
CB2 test	0.931	0.914	0.906	0.908	0.918	0.920	0.904	0.927	0.807	0.917	0.891
CB10/CB1A test	0.923	0.940	0.769	0.800	0.887	0.925	0.915	0.979	0.822	0.924	0.915
MACCS											
CB1 test	0.892	0.902	0.882	0.902	0.907	0.893	0.848	0.880	0.796	0.832	0.903
CB2 test	0.917	0.924	0.928	0.923	0.920	0.915	0.891	0.895	0.839	0.872	0.928
CB10/CB1A test	0.935	0.970	0.969	0.955	0.945	0.948	0.868	0.970	0.834	0.813	0.937
ECFP6											
CB1 test	0.861	0.916	0.896	0.899	0.909	0.917	0.893	0.899	0.764	0.942	0.912
CB2 test	0.936	0.945	0.940	0.947	0.930	0.934	0.900	0.926	0.811	0.957	0.953
CB10/CB1A test	0.939	0.979	0.979	0.984	0.982	0.944	0.873	0.972	0.872	0.973	0.978

^aEach bold entry shows the highest metric value among the machine learning models using different algorithms.

Table 4.

Ranked Molecular Descriptor-Based Prediction Scores for Each Machine Learning Algorithm by Metrics (Average over Three Datasets)^a

algorithms	AUC	F1_score	ACC	Cohen's κ	MCC	precision	recall	mean	rank
SVM	0.926	0.573	0.900	0.504	0.540	0.495	0.802	0.677	2
MLP_1	0.902	0.560	0.902	0.503	0.542	0.476	0.813	0.671	3
MLP_2	0.843	0.517	0.856	0.390	0.407	0.466	0.639	0.588	11
MLP_3	0.858	0.549	0.876	0.448	0.468	0.534	0.616	0.621	9
MLP_4	0.901	0.566	0.887	0.454	0.480	0.472	0.735	0.642	7
MLP_5	0.912	0.570	0.906	0.499	0.522	0.494	0.738	0.663	6
RF	0.911	0.582	0.907	0.503	0.521	0.503	0.728	0.665	5
ABDT	0.940	0.595	0.920	0.556	0.602	0.504	0.875	0.713	1
DT	0.816	0.562	0.901	0.478	0.492	0.497	0.682	0.632	8
NB	0.925	0.561	0.878	0.470	0.526	0.451	0.887	0.671	3
logistic	0.877	0.505	0.842	0.415	0.464	0.423	0.821	0.621	9

^a Each bold entry shows the highest metric value among the machine learning models using different algorithms.

Table 5.

Ranked MACCS Fingerprint-Based Prediction Scores for Each Machine Learning Algorithm by Metrics (Average over Three Datasets)^a

algorithms	AUC	F1_score	ACC	Cohen's κ	MCC	precision	recall	mean	rank
SVM	0.915	0.526	0.847	0.441	0.498	0.438	0.878	0.649	7
MLP_1	0.932	0.553	0.884	0.492	0.545	0.480	0.860	0.678	1
MLP_2	0.923	0.524	0.854	0.452	0.520	0.431	0.920	0.661	4
MLP_3	0.927	0.561	0.900	0.489	0.521	0.475	0.775	0.664	3
MLP_4	0.924	0.555	0.892	0.482	0.524	0.465	0.817	0.666	2
MLP_5	0.919	0.531	0.869	0.441	0.496	0.430	0.850	0.648	8
RF	0.869	0.485	0.831	0.374	0.431	0.392	0.818	0.600	10
ABDT	0.915	0.548	0.907	0.491	0.525	0.465	0.766	0.660	6
DT	0.823	0.520	0.889	0.435	0.455	0.454	0.675	0.607	9
NB	0.839	0.459	0.792	0.326	0.388	0.368	0.817	0.570	11
logistic	0.923	0.525	0.852	0.445	0.518	0.423	0.938	0.661	4

^a Each bold entry shows the highest metric value among the machine learning models using different algorithms.

Table 6.

Ranked ECFP6 Fingerprint-Based Prediction Scores for Each Machine Learning Algorithm by Metrics (Average over Three Datasets)^a

algorithms	AUC	F1_score	ACC	Cohen's κ	MCC	precision	recall	mean	rank
SVM	0.912	0.572	0.909	0.511	0.544	0.479	0.793	0.674	9
MLP_1	0.947	0.619	0.922	0.568	0.601	0.534	0.838	0.719	3
MLP_2	0.938	0.609	0.920	0.552	0.582	0.516	0.818	0.705	5
MLP_3	0.943	0.610	0.923	0.564	0.598	0.530	0.828	0.714	4
MLP_4	0.940	0.605	0.922	0.554	0.582	0.542	0.782	0.704	6
MLP_5	0.932	0.600	0.915	0.544	0.583	0.505	0.848	0.704	6
RF	0.889	0.575	0.902	0.479	0.492	0.501	0.683	0.646	10
ABDT	0.932	0.573	0.907	0.520	0.560	0.489	0.827	0.687	8
DT	0.816	0.487	0.878	0.398	0.422	0.416	0.659	0.582	11
NB	0.957	0.626	0.923	0.572	0.603	0.536	0.838	0.722	2
logistic	0.948	0.624	0.924	0.574	0.609	0.527	0.856	0.723	1

^a Each bold entry shows the highest metric value among the machine learning models using different algorithms.

Table 7. Top 10 Molecular Descriptor Features for the ABDT Classification of Orthosteric and Allosteric Ligands on the CB1 Receptor

features	min	max	mean	SD	skewness	kurtosis
smr_VSA7	CB1O	11.760	147.629	64.037	25.234	0.372
	CB1A	0.000	168.852	56.255	23.421	0.692
slogp_VSA2	CB1O	0.000	82.679	29.892	15.362	0.737
	CB1A	0.000	94.932	42.720	17.653	0.187
slogp_VSA5	CB1O	0.000	122.499	48.751	24.375	0.458
	CB1A	0.000	96.815	34.469	18.617	0.303
Chi1v	CB1O	5.520	20.013	11.224	1.826	0.727
	CB1A	3.813	18.165	10.420	2.782	0.610
slogp_VSA3	CB1O	0.000	34.435	9.769	7.555	0.921
	CB1A	0.000	32.396	13.916	7.599	-0.067
peoe_VSA1	CB1O	0.000	30.531	9.963	5.314	0.473
	CB1A	0.000	29.744	14.913	5.586	-0.007
Chi3v	CB1O	3.196	11.702	6.511	1.547	0.715
	CB1A	1.862	13.005	5.771	1.964	1.173
smr_VSA3	CB1O	0.000	30.001	8.879	6.819	0.550
	CB1A	0.000	35.936	14.548	9.091	0.222
slogp_VSA8	CB1O	0.000	28.333	8.584	8.689	0.521
	CB1A	0.000	22.973	10.330	6.172	-0.169
peoe_VSA9	CB1O	0.000	46.264	12.496	9.333	0.585
	CB1A	0.000	36.107	12.150	8.637	0.545

Table 8.

Top 10 MACCS Fingerprint Features for the Logistic Regression Classification of Orthosteric and Allosteric Ligands on CB1 Receptor

MACCS index	Number of CB1 orthosteric ligands having this feature	Number of CB1 allosteric ligands having this feature	Example structure	Representative compounds with the features	Category	CAS Registry Number
144	3 (0.8%)	19 (9.1%)		CB1 allosteric ligand	CB1 allosteric ligand	1207203-33-5
76	146 (38.8%)	161 (77.4%)		CB1 allosteric ligand	CB1 allosteric ligand	1160157-67-4
130	22 (5.8%)	50 (24.0%)		CB1 allosteric ligand	CB1 allosteric ligand	1626414-43-4
84	172 (45.7%)	164 (78.8%)		CB1 allosteric ligand	CB1 allosteric ligand	1377838-06-6
115	80 (21.3%)	27 (13.0%)		CB1 orthosteric ligand	CB1 orthosteric ligand	942124-70-1
78	109 (29.0%)	12 (5.8%)		CB1 orthosteric ligand	CB1 orthosteric ligand	903889-18-9
9	322 (85.6%)	191 (91.8%)		CB1 allosteric ligand	CB1 allosteric ligand	1207203-44-8

MACCS index	Number of CB1 orthosteric ligands having this feature	Number of CB1 allosteric ligands having this feature	Representative compounds with the features		
			Example structure	Category	CAS Registry Number
81	126 (33.5%)	159 (76.4%)		CB1 allosteric ligand	1207203-74-4
145	34 (9.0%)	4 (1.9%)		CB1 orthosteric ligand	1034925-58-0
13	274 (72.9%)	202 (97.1%)		CB1 allosteric ligand	1626414-47-8