Journal of Medicinal Chemistry

ACS AUTHORCHOICE

Article

Cite This: J. Med. Chem. 2020, 63, 8761−8777

pubs.acs.org/jmc

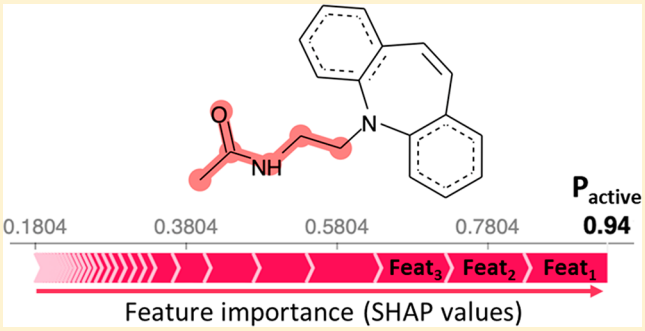
Interpretation of Compound Activity Predictions from Complex Machine Learning Models Using Local Approximations and Shapley Values

Raquel Rodríguez-Pérez[†,‡] and Jürgen Bajorath*[,†]

[†]Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Endenicher Allee 19c, D-53115 Bonn, Germany

[‡]Department of Medicinal Chemistry, Boehringer Ingelheim Pharma GmbH & Co. KG, Birkendorfer Straße 65, 88397 Biberach an der Riß, Germany

ABSTRACT: In qualitative or quantitative studies of structure−activity relationships (SARs), machine learning (ML) models are trained to recognize structural patterns that differentiate between active and inactive compounds. Understanding model decisions is challenging but of critical importance to guide compound design. Moreover, the interpretation of ML results provides an additional level of model validation based on expert knowledge. A number of complex ML approaches, especially deep learning (DL) architectures, have distinctive black-box character. Herein, a locally interpretable explanatory method termed Shapley additive explanations (SHAP) is introduced for rationalizing activity predictions of any ML algorithm, regardless of its complexity. Models resulting from random forest (RF), nonlinear support vector machine (SVM), and deep neural network (DNN) learning are interpreted, and structural patterns determining the predicted probability of activity are identified and mapped onto test compounds. The results indicate that SHAP has high potential for rationalizing predictions of complex ML models.



INTRODUCTION

Compound bioactivity prediction and structure−activity relationship (SAR) analysis are major applications of machine learning (ML) in pharmaceutical research.[1−6] Supervised ML methods are trained to search for structural patterns that differentiate between active and inactive compounds. Since prospective predictions using such activity models provide decision support and guidance for compound exploration and design, there is a high level of interest in obtaining accurate models and in rationalizing their predictions.[7−9] However, while much attention has been paid to improving the predictive performance of ML models, interpreting the predictions currently is an underinvestigated area, despite its high relevance.[10,11]

While statistical performance measures and method validation procedures are of critical importance for ML, they do not provide scientific insights into predictions, which can typically only be achieved on the basis of expert knowledge. On the other hand, rationalizing model decisions would assign priority to meaningful predictions, help to extract knowledge from ML models, and also increase the acceptance of and confidence in predictions in pharmaceutical research.[5,12,13] In activity prediction, model interpretation generally relies on the identification of chemical features that determine predictions.[14,15] For simplistic linear (Q)SAR models, the interpretation of structural and/or property changes that modulate activity is often straightforward.[13] However, the situation fundamentally changes when ML models become complex, which often increases predictive performance at the expense of interpretability, ultimately leading to the frequently quoted "black-box" character of ML model and their predictions.[13,15] For example, the random forest (RF)[16] and support vector machine (SVM)[17] algorithms are robust and well-performing ML methods that have become very popular in the field. However, RF and SVM models are very difficult to interpret and exhibit black-box character, for different reasons. In the case of RF, this is largely due to the generation of large decision tree ensembles, leading to statistically driven decisions; in the case of SVM, black-box character results from the use of nonlinear kernels to facilitate data mapping into feature reference spaces of increasing dimensionality.[18]

Currently, compound activity data grow at unprecedented rates,[19,20] leading to emerging big data phenomena in medicinal chemistry[19] and catalyzing the application of deep learning (DL)[21] strategies for activity prediction. Among ML methods, DL architectures have shown particular promise in data-rich fields such as image analysis[22] or natural language

Special Issue: Artificial Intelligence in Drug Discovery

Received: July 8, 2019
Published: September 12, 2019

DOI: 10.1021/acs.jmedchem.9b01101
J. Med. Chem. 2020, 63, 8761−8777

processing[23] and deep neural networks (DNNs) also gain increasing popularity in chemical informatics and drug design.[24−26] Although some successful applications in compound design and activity prediction using DNNs have been reported, it remains unclear at present whether DL might provide a consistent advantage over other ML methods in at least some application scenarios.[27−31] However, DNNs have higher complexity than other ML models and their black-box character is notorious. Any form of model diagnostics becomes essentially prohibitive for DNNs, and domain experts struggle to understand why DNN models succeed or fail,[32] which hinders advances in the field.

Several interpretation strategies have been proposed to reduce the black-box nature of ML models.[13] These approaches can essentially be divided into model-specific and model-agnostic (or model-independent) strategies. As a model-specific approach, feature weighting has been applied to better understand predictions of SVM[18,33] and RF models.[34] As a model-agnostic method, sensitivity analysis can be used to investigate the influence of systematic feature value changes on the model output.[35] Sensitivity analysis has been applied to different ML algorithms including neural networks[36] but becomes quickly inefficient with increasing dimensionality of models and has thus hardly been used in chemical informatics.[13] An exception is provided by investigating partial derivatives as a form of local sensitivity analysis that has been applied in QSAR modeling.[13] Here, for a given compound, a perturbation is introduced to an individual feature and calculation of the partial derivative provides an estimation of its contribution to model performance.[37,38] However, effective use of partial derivatives is also limited given its intrinsic focus on individual features. A principal advantage of model-agnostic over model-specific interpretation approaches, if they can be established, is that model-agnostic analysis alleviates the need to balance model performance and interpretability.[39,40]

In this work, we introduce a conceptual new agnostic interpretation method for ML models of arbitrary complexity used for activity prediction. The Shapley additive explanations (SHAP) approach[41] is an extension of local interpretable model-agnostic explanations (LIME)[42] according to which feature weights are represented as Shapley values from game theory.[43] As shown herein, SHAP is capable of interpreting activity predictions from complex ML models. Features that increase or reduce the probability of predicted activity are identified and mapped onto molecular graphs to identify and visualize structural patterns that determine predictions.

RESULTS

Principles of Explanation Models and the LIME Approach. *Explanation Model.* The principal goal of an explanation model g is to simplify or locally approximate a complex model f that cannot be directly interpreted. Additive feature attribution methods generate an explanation model via a linear function of binary variables, as shown in eq 1:

$$g(x') = \phi_0 + \sum_{i=1}^{M} \phi_i x'_i \tag{1}$$

where $x' \in \{0,1\}^M$, M is the number of input features, and $\phi_i \in \mathbb{R}$.[42] The presence or absence of a feature value impacts the model, which can be referred to as a feature contribution (ϕ_i). Accordingly, a weight must be assigned to each variable. Therefore, the SHAP method has been devised, which represents an extension of the LIME approach.

LIME. The LIME methodology generates the explanation ξ of an instance x according to eq 2:

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}}\ \mathcal{L}(f, g, \pi_x) + \Omega(g) \tag{2}$$

where G is a class of interpretable (linear) models, $\mathcal{L}$ is the loss function to minimize, π_x the proximity measure between an instance z and x (kernel defining locality), and $\Omega(g)$ an optional regularization term to control (limit) model complexity.[42]

For the interpretation of a given test instance x, the following procedure is applied.

(i) Artificial samples are obtained by permuting features of the test instance x.
(ii) These samples are weighted by the value of a kernel calculated for them and x.
(iii) A model g is trained to predict $f(x)$ with coefficients corresponding to feature importance estimates.

It follows that LIME builds a linear model g in a feature region proximal to the test instance, although model f might be nonlinear, as illustrated in Figure 1. This figure also shows that

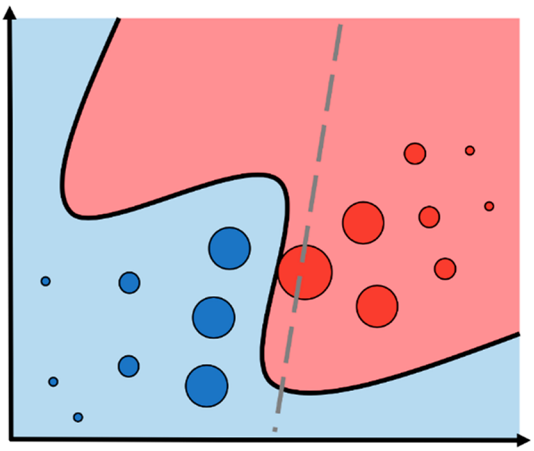


Figure 1. Local approximations for model interpretation. The active (red) and inactive (blue) regions in feature space correspond to the decision function of the complex model f. The dashed gray line represents the decision function of the simple explanation model g, which locally approximates the global model. The largest red dot is the active instance x to be explained, while the other dots are artificial samples that are weighted by the kernel function with respect to x.

samples similar to x receive high weights, due to the application of the kernel function. This conceptual framework provides the basis for the development of the SHAP methodology detailed in the following.

SHAP Method. *Shapley Value Concept.* Shapley values from cooperative game theory provide a connection between LIME and the SHAP methodology. Specifically, Shapley values were introduced in the 1950s to measure contributions of individual players to a collaborative game.[43] They provide a theoretically grounded partition of payoff or credit among members of a team by considering the average of all contributions made by a player.[43] This concept can be applied to feature attributions by considering the success of a team (or total credit) as an output (prediction), and each player's contribution (or player's payoff) as the feature importance.

DOI: 10.1021/acs.jmedchem.9b01101
J. Med. Chem. 2020, 63, 8761−8777

Therefore, in this context, Shapley values facilitate the distribution of a model's prediction resulting from an input feature vector over the individual features.

To obtain the contribution of a feature i, all operations by which a feature might have been added to the set ($N!$) and a summation over all possible sets (S) is considered. For any feature sequence, the marginal contribution through addition of feature i is given by $[f(S\cup\{i\}) - f(S)]$. The resulting quantity is weighted by the different possibilities the set could have been formed prior to feature i's addition ($|S|!$) and the remaining features could have been added ($(|N| - |S| - 1)!$). Hence, the importance of a given feature i is defined by eq 3:

$$\phi_i = \frac{1}{N!} \sum_{S\subseteq N\setminus\{i\}} |S|!(|N| - |S| - 1)![f(S \cup \{i\}) - f(S)] \quad (3)$$

It follows that Shapley values represent a unique way to divide a model's output among feature contributions satisfying three axioms: *local accuracy* (or additivity), *consistency* (or symmetry), and *nonexistence* (or null effect).

SHAP Formalism. Additive feature attribution methods typically do not consider two properties that are of high relevance for assessing feature importance, i.e., *local accuracy* and *consistency*, as referred to above. Taking these axiomatic properties into account was a main motivation for proposing the SHAP concept.[41] The property *local accuracy* forces the sum of individual feature attributions to be equal to the original model prediction. In addition, *consistency* ensures that feature importance correctly accounts for different models on a relative scale. Hence, if a change in a feature value has larger impact on a model A than a model B, feature importance should be larger in A. These properties can be considered by expressing feature weights as Shapley values.[43]

A weighting procedure for artificial samples is a key aspect for connecting Shapley values to the LIME approach, which allows the approximation of Shapley values. In LIME, heuristic choices are made to select $\mathcal{L}$, $\Omega(g)$, and π_x. By contrast, the SHAP method introduces a special kernel function that is related to the Shapley value definition, assuming that feature weights follow the two axioms of interpretability.[41] Specifically, SHAP uses the following procedure for interpreting an instance x:

(i) Training data is organized by k-means clustering and the k samples are weighted by the number of training instances they represent. These samples constitute a background data set with "typical" feature values.
(ii) Artificial samples are obtained by replacing features of the test instance x with the values from the background data set.
(iii) These artificial samples are weighted by the value of the SHAP kernel calculated for them and x.
(iv) A weighted linear regression model g is trained to predict $f(x)$. The model coefficients are Shapley values corresponding to feature importance estimates.

Sampling all possible feature subsets is time-consuming. Therefore, the input vector is permutated for an individual prediction by setting its features on and off, thereby examining feature influence. Herein, 1000 artificial samples were generated in each case and missing features were simulated by replacing them with the values obtained from a k-means clustering of the training set (k = 100). A feature obtained a large weight if its replacement with an artificial (non-informative) value led to a significant change in model output. Weights of artificial samples were determined according to the number of feature-addition sequences that a given subset accounted for on the basis of the SHAP kernel. Local linear regression resulted in coefficients representing feature weights as Shapley values. These weights indicate how important a feature is for a given prediction and include the direction (sign) of feature influence. The expected explanatory value is calculated as the mean of the model output probability over training set instances. For a given compound, the original output probability (of activity) given by model f is then retrieved by summing the expected (or base) value and all SHAP values.

Model Building and Analysis Strategy. ML models were built for 10 activity classes summarize in Table 1. These classes were assembled on the basis of specific structural and activity data selection criteria detailed in the Experimental Section. As negative training and test instances, compounds with unknown activity status were considered inactive and randomly assembled, as also reported in the Experimental Section. Feature contributions were systematically calculated for test set compounds. First, model performance for three different ML algorithms and two molecular representations is reported. Then, the effect of feature removal is investigated. SHAP results for RF models are compared to Gini importance, and the relationship between SHAP values obtained for different ML methods is examined. Next, representative examples are shown to illustrate SHAP results. Individual predictions using ML algorithms are interpreted and differences in feature importance are explored. Furthermore, for individual predictions, important (fingerprint) features are mapped onto compounds and visualized.

Table 1. Compound Data Sets[a]

CHEMBL identifier	target	no. compounds	no. ASs	mean pK_i
229	α-1a adrenergic receptor	243	80	7.8
4860	Apoptosis regulator Bcl-2	283	67	9.0
244	Coagulation factor X	679	154	7.5
264	Histamine H3 receptor	955	216	8.0
237	κ opioid receptor	716	160	7.5
344	Melanin-concentrating hormone receptor 1	409	73	7.4
259	Melanocortin receptor 4	443	57	6.9
1946	Melatonin receptor 1B	285	70	8.2
233	μ opioid receptor	831	194	7.6
4792	Orexin receptor 2	399	81	6.9

[a]Reported are the CHEMBL identifier, target name, number of compounds, number of analog series (ASs), and mean pK_i values for 10 compound activity classes.

While our study is focused on method development and evaluation, it is essential to carry out the analysis on newly generated ML models and their predictions to ensure independence of ML assessment (rather than reliance on previously reported models) and reproducibility of the results.

Global Model Performance. Accurate predictions are a key requirement for meaningful model interpretation. If ML models are not predictive, the prioritized chemical patterns do not correlate well with activity prediction. Thus, initially, the predictive performance of SVM, RF, and DNN models over different compound activity classes was determined. Models were built on the basis of the state-of-art ECFP4 and easy-to-

DOI: 10.1021/acs.jmedchem.9b01101
J. Med. Chem. 2020, 63, 8761−8777

Table 2. Classification Performance[a]

	ECFP4			MACCS		
metric	SVM	RF	DNN	SVM	RF	DNN
AUC	0.98 (0.02)	0.98 (0.02)	0.98 (0.02)	0.97 (0.02)	0.97 (0.02)	0.97 (0.02)
BA	0.89 (0.30)	0.84 (0.05)	0.91 (0.03)	0.88 (0.04)	0.84 (0.04)	0.89 (0.03)
MCC	0.87 (0.03)	0.80 (0.07)	0.88 (0.03)	0.83 (0.06)	0.79 (0.06)	0.81 (0.05)

[a]Area under the ROC curve (AUC), balanced accuracy (BA), and Matthew's correlation coefficient (MCC). Mean (and standard deviation) values are reported across 10 activity classes. Performance values are given for two molecular representations (ECFP4 and MACCS) and three ML methods (SVM, RF, and DNN).

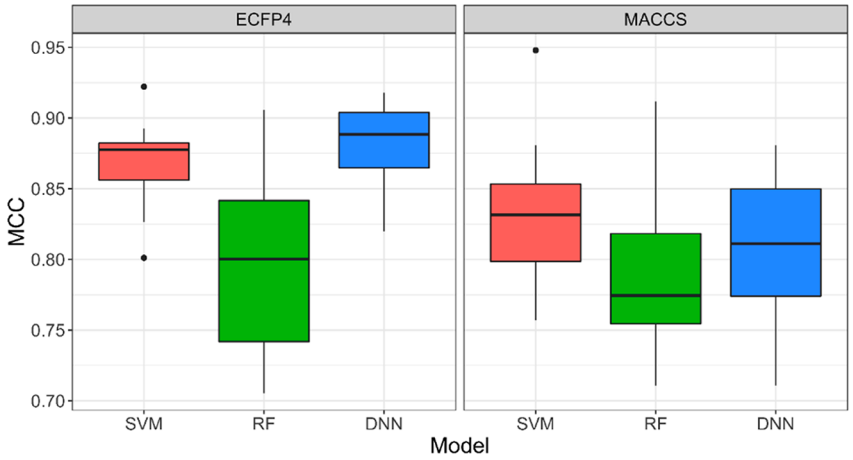


Figure 2. Global classification performance. Boxplots show value distributions of Matthew's correlation coefficient (MCC) across 10 compound data sets using SVM (red), RF (green), DNN (blue) models and two fingerprints (ECFP4, left; MACCS, right).

understand MACCS fingerprints. Table 2 reports average model performance on the basis of the AUC, BA, and MCC measures. Overall, activity predictions for the 10 activity classes were consistently accurate for the investigated methods and molecular representations, hence providing a sound basis for model analysis. Overall, rankings of test compounds yielded AUC values greater than 0.9, BA of ∼90%, and MCC values of around 0.8 or larger. Figure 2 reports the distribution of MCC values for all ML method/representation combinations. As anticipated, MCC values were larger for ECFP4 than MACCS, albeit by a confined margin. In addition, RF predictions were generally slightly less accurate than SVM and DNN predictions. Although hyperparameter combinations were optimized (see Experimental Section), alternative parameter settings did not have a large influence on the predictions because active compounds were overall easily distinguishable from random ZINC examples. Taken together, the results showed that the test system setup was appropriate for our proof-of-principle investigation of a new model interpretation methodology.

Feature Importance. To interpret the prediction for a test compound, SHAP calculations were carried out resulting in a set of feature weights. Initially, the distributions of ECFP4 features with nonzero SHAP values (feature weights) were determined for all test compounds. Figure 3a shows how many feature variables were contributing to the RF, SVM, and DNN predictions of individual test instances. SVM and DNN distributions were centered on smaller values than RF, indicating that more features were required to provide local explanations for RF predictions. The average number of features with nonzero SHAP values for a test instance was 68 and 67 for SVM and DNN, respectively, and 96 for RF. These numbers represented less than 10% of the entire ECFP feature population obtained for the activity classes, revealing that limited numbers of features were important for the predictions.

Because some features with nonzero SHAP values might not contribute significantly to predictions, absolute SHAP values of features were normalized with respect to the total sum of SHAP values for a given prediction, resulting in a percentage value for a feature. This percentage represents the fraction of feature weights that a given variable is accounting for, considering both negative and positive contributions. Thus, the cumulative SHAP percentage for a given number of top-ranked features can be calculated per test instance. Figure 3b shows the distributions of cumulative SHAP percentage values for different numbers of top-ranked features. The distributions were nearly identical for all three ML methods and showed that the top-1, -5, -10, and -20 ranked features generally corresponded to 7%, 25%, 40%, and 60% of the cumulative (total) feature weights of a prediction, respectively. These findings indicated that top-ranked features provided sufficient information for model interpretation.

Feature Elimination. The next step was exploring whether SHAP values indeed identified features that were important for predictive performance. Therefore, for each data set and ML model, SHAP values were calculated for all test compounds. Then, absolute SHAP values were averaged over test compounds to obtain an ECFP4 feature importance ranking. Finally, features were systematically eliminated, either randomly or in the order of SHAP ranking, and the ML models were generated again using the reduced feature sets. Following this protocol, RF, SVM, and DNN control models were built after removal of 4, 10, 80, 160, and 320 ECFP4 features. Figure 4 shows the median MCC values of SVM, RF, and DNN models across all activity classes for varying numbers of features. The results revealed that random elimination of up

DOI: 10.1021/acs.jmedchem.9b01101
J. Med. Chem. 2020, 63, 8761−8777

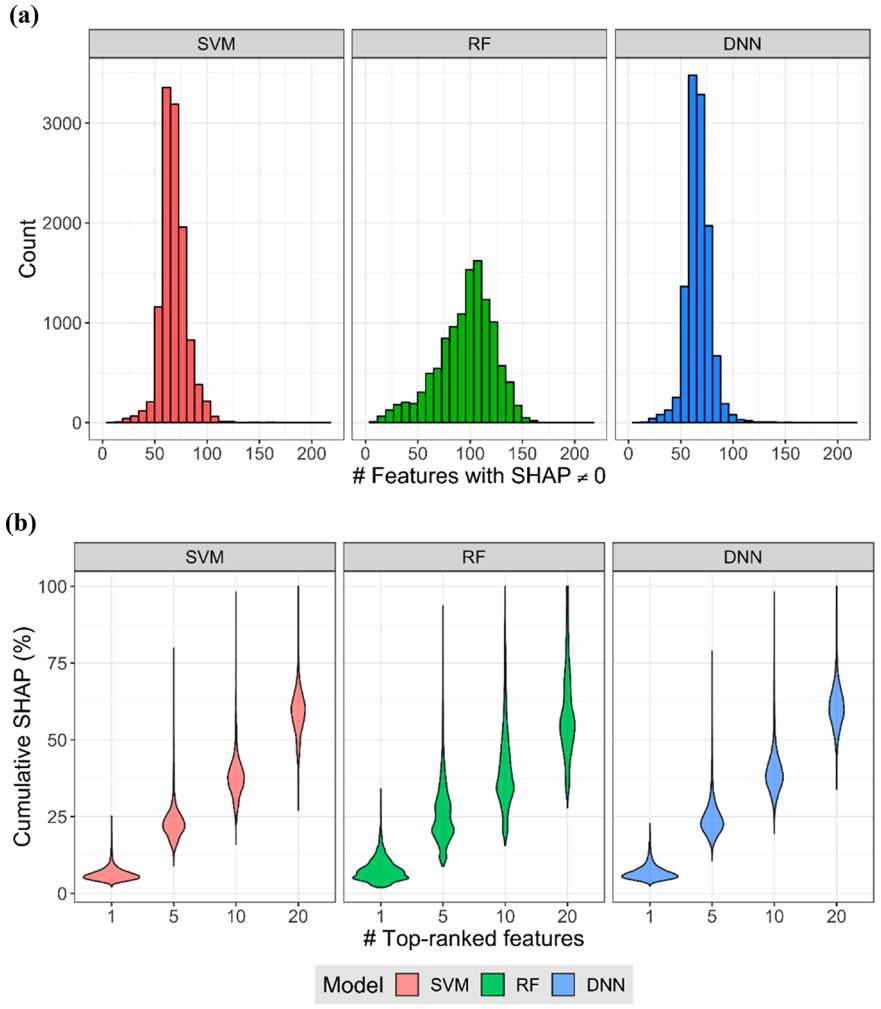


Figure 3. Distribution of SHAP values. (a) shows distributions of features with nonzero SHAP values over all test compounds (Count) for RF, SVM, and DNN predictions. (b) shows distributions of cumulative SHAP percentage values for different numbers of top-ranked features.

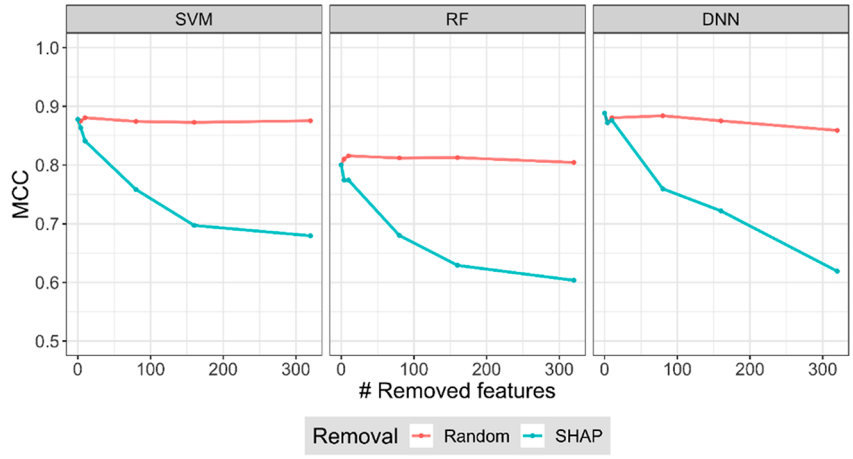


Figure 4. Feature removal. MCC values are shown for varying numbers of ECFP4 features, which were removed randomly (red) or according to decreasing mean absolute SHAP values (blue). Results are shown for SVM (left), RF (center), and DNN (right) models.

to 320 features did not notably affect the performance of ML models, which remained essentially constant, providing further evidence for general ECFP4 feature redundancy. By contrast, removal of features with large average SHAP values led to a substantial decrease in model performance for the three ML algorithms.

For all ML methods, the MCC value distribution after feature removal according to SHAP values was significantly larger than the one after random elimination (Wilcoxon test, p-values $\ll 0.0001$). These results confirmed that SHAP values provided a quantitative measure of feature importance for predictions using different ML models.

DOI: 10.1021/acs.jmedchem.9b01101
J. Med. Chem. 2020, 63, 8761−8777

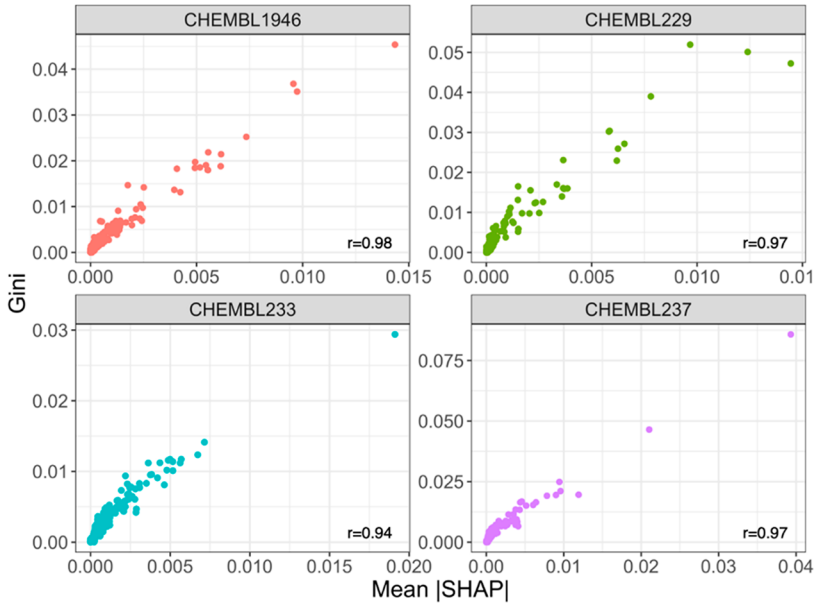


Figure 5. Relationship between SHAP and Gini importance. For four activity classes, RF models were built to predict the activity of test compounds. For each ECFP4 feature, mean absolute SHAP values for test compounds and Gini importance are reported for RF models. In addition, the correlation coefficient for feature weighting methods is reported.

SHAP versus Gini Importance. In an additional control calculation, SHAP feature weights were compared to Gini importance,[34] which has become a popular measure for the assessment of variable importance in decision tree-based methods such as RF. Gini importance is equivalent to the mean decrease in Gini "impurity", which measures the probability of a new sample to be incorrectly classified at a given node in a tree weighted by the proportion of samples representing the data partition. Gini feature importance values were calculated during RF model building[56] and were thus not dependent on test instances. Gini calculations yielded absolute (nonsigned) values, which were thus compared to mean absolute SHAP values determined from predictions of all test compounds. Figure 5 compares feature weights obtained using both approaches for RF models of four activity classes. Each point represents the weights for a given feature using SHAP and Gini importance. There was strong correlation between these orthogonal feature weights (i.e., one derived on the basis of training, the other on the basis of testing), without any outlier or notable inconsistency. However, while Gini feature importance is confined to decision tree methods, SHAP is generally applicable.

SHAP Comparison. Next, relationships between SHAP values for the same compound sets and different ML methods were examined. Despite algorithmic differences, which might affect variable prioritization, ML models with predictive power should detect similar chemical patterns that differentiate between active and inactive compounds for a given molecular representation.

Figure 6 shows mean absolute SHAP values for test compounds from two activity classes. SHAP values originating from SVM and RF models were compared in a pairwise manner to corresponding values from DNN models based upon the ECFP4 (Figure 6a) and MACCS (Figure 6b) representations. Correlation coefficients were high, ranging from 0.90 to 0.98 for ECFP4 and from 0.83 to 0.95 for MACCS. Highly weighted features were consistently prioritized for models generated with all ML methods, thus confirming algorithm-independent consistency of feature relevance. We note that features that are important in a local explanation model might not be globally relevant. Therefore, some features influencing individual predictions might yield small (but nonzero) mean SHAP values because they were not prioritized in the majority of explanation models.

Feature weight relationships between different ML methods were also examined across all activity classes. Therefore, correlation between mean absolute SHAP values for models generated with different methods was determined. The resulting distributions or correlation values are shown in Figure 7. All method combinations displayed high correlation of feature importance, especially SVM and DNN, with a median correlation coefficient of 0.97 and 0.95 for ECFP4 and MACCS, respectively. SHAP mean values were overall more strongly correlated for ECFP4- than MACCS-based models.

Taken together, the results in Figures 6 and 7 revealed that SHAP values of features originating from models built using ML algorithms were highly correlated, showing that the different methods prioritized similar chemical patterns for predictions that were consistently detected in the basis of SHAP values.

Visualization of SHAP Values. To rationalize model predictions, features with highest SHAP values for individual predictions were extracted, first for the simplistic MACCS fingerprint that encodes the presence or absence of predefined chemical patterns. Figure 8 shows MACCS feature weights for the correct prediction of three compounds using SVM, RF, and DNN models. The first compound (Figure 8a) was an antiapoptotic Bcl-2 inhibitor and the second (Figure 8b) a melanin-concentrating hormone receptor 1 antagonist. The third compound (Figure 8c) was a factor X inhibitor. For each ML model and test compound, SHAP values for MACCS features are reported in a separate graph. Positive and negative feature contributions are identified using sequential red and blue arrows, respectively. The length of each arrow is proportional to the SHAP value for a given feature, and the MACCS keys corresponding to the top-ranked variables with

DOI: 10.1021/acs.jmedchem.9b01101
J. Med. Chem. 2020, 63, 8761−8777

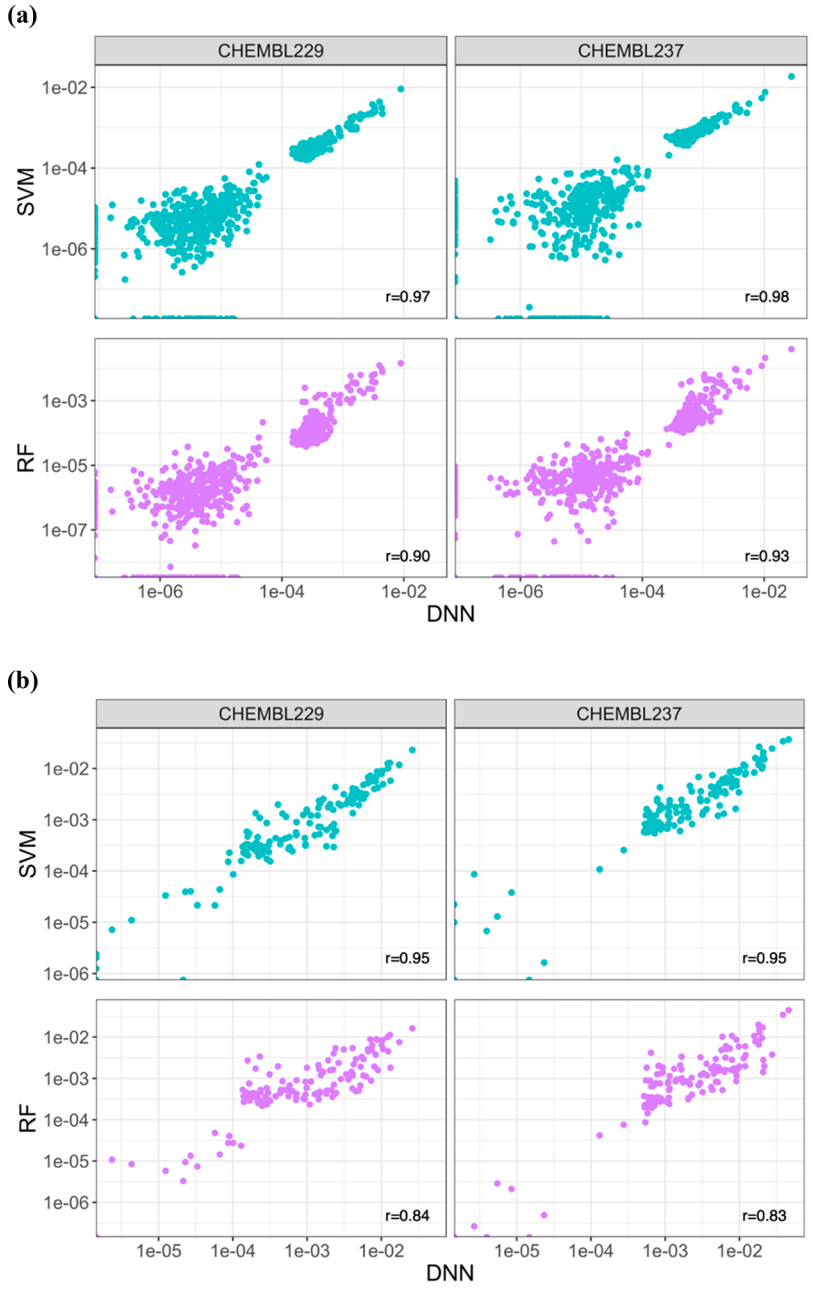


Figure 6. Comparison of SHAP values. Mean absolute SHAP values for features originating from different ML models are compared in a pairwise manner. Each data point represents a pairwise value comparison for a given feature. Different ML models were generated on the basis of (a) ECFP4 and (b) MACCS.

(largest absolute SHAP value) are given. The expected value is obtained as the average model output over training set instances and corresponds to the predicted probability of a test compound with unknown feature values. It is also referred to as the base probability. SHAP values quantify the influence of a given feature on a prediction and modify the expected value. When SHAP values are added to this base value, the output probability of the original ML model is obtained (shown in bold).

The three compounds in Figure 8 were correctly classified as actives by the three ML algorithms. Moreover, different methods shared most top-ranked features, indicating that similar chemical patterns determined the prediction of a given compound. However, the absolute importance values differed between ML methods and other features with smaller SHAP values also contributed to the predictions, resulting in different final output probabilities. DNN models produced the highest output probabilities of activity for these test compounds, whereas RF models gave the smallest ones.

Figure 9 shows feature weights for three other exemplary compounds, which were represented by ECFP4, including a κ opioid receptor (Figure 9a), melanocortin receptor 1B (Figure 9b), and orexin receptor 2 (Figure 9c) ligands. ML models correctly predicted these active compounds and SHAP values were calculated to examine the prioritized features for the predictions. In this case, positive and negative feature contributions were displayed using the ECFP4 feature index (obtained after fingerprint folding). Again, most highly weighted features were common to SVM, RF, and DNN models. RF gave the overall lowest output probability, due to

DOI: 10.1021/acs.jmedchem.9b01101
J. Med. Chem. 2020, 63, 8761−8777

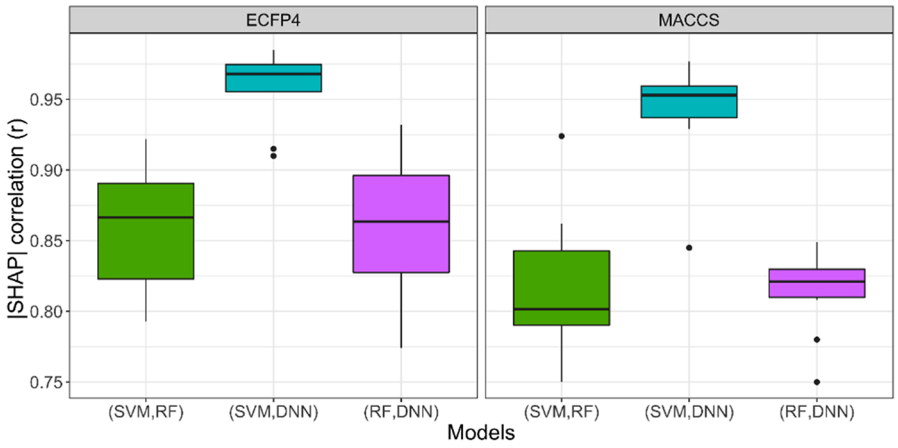


Figure 7. SHAP value correlation between ML models. For each activity class, the correlation between mean absolute SHAP values was calculated for different ML models. Boxplots report the pairwise correlation of SVM vs RF (green), SVM vs DNN (blue), and RF vs DNN (purple) for two molecular representations (ECFP4, left; MACCS, right).

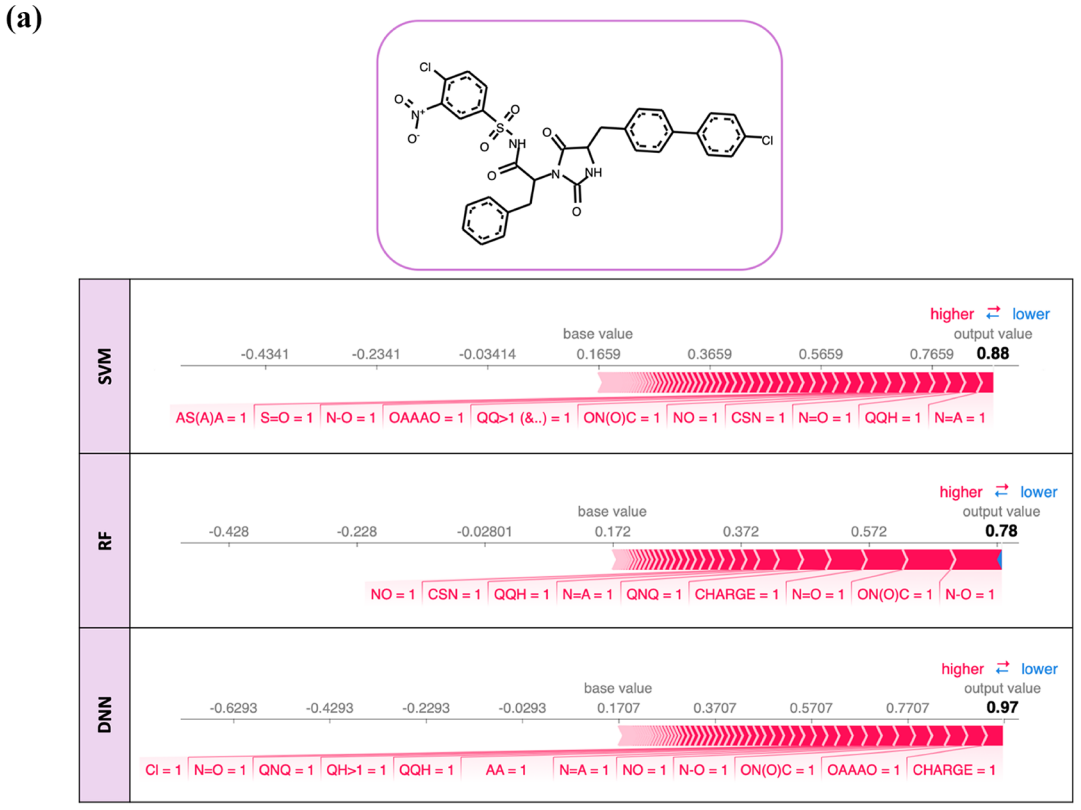
smaller individual feature weights. The corresponding substructures of top-ranked important features shared by the three algorithms were mapped onto the compounds. In Figure 9a, feature with index #566 is highlighted, which was relevant for the three models applying a SHAP threshold value of ∼0.07. Feature #566 was ranked top-1 (SVM), -2 (RF), and -2 (DNN). On the other hand, in Figure 9b, the highlighted feature #637 was ranked top-5 (SVM), -1 (RF), and -6 (DNN). However, in the latter case, features ranked higher than #637 (SMILES, [CH2]CNC(C)=O) represented substructures of #637 (#1010, CC; #29, [CH2]NC; #118, [CH2]NC(C)=O; #236, CC([NH])=O; #960, [CH2]C-[NH]) and were thus correlated. Finally, in Figure 9c, highlighted features include #843 (ranked top-1 (SVM), -1 (RF), and -2 (DNN)) and #268 (ranked top-2 (SVM), -3 (RF), and -1 (DNN)). The SHAP values representations in Figures 8 and 9 provide global explanations for a given prediction and enable comparison of feature importance across different models and methods.

Feature Mapping onto Compounds. The use of the ECFP4 fingerprint made it possible to map highly weighted topological features onto molecular graphs and analyze resulting substructures. Although ECFP4 folding might lead to feature "collisions" (i.e., different atom environments might be encoded by the same element), such collisions were only very rarely observed for individual compounds because of their generally low number of hash values compared to the size of the folded fingerprint. In global model interpretation, a unique weight is obtained for each feature. SHAP values explain individual predictions, and for a given compound, correspondence between a given feature and substructure is generally unequivocal. Furthermore, different mapped features might contribute to the formation of coherent, overlapping, or distinct substructures. Figure 10 provides an example for the rationalization of a prediction on the basis of SHAP values. Figure 10a depicts the mapping of the most relevant features onto a compound active against the κ opioid receptor, and Figure 10b gives an overview of the positive and negative feature contributions. All three ML models correctly predicted this test compound, and the substructures resulting from mapping of features that determined these predictions were explored. For feature mapping, a threshold should be defined that can be based on the absolute SHAP value, the signed value (accounting for positive or negative contributions), or the number of top-ranked features. Therefore, depending on the application, different types of threshold values can be used. In this case, the threshold was iteratively varied, and results for different SHAP threshold values are shown in the figure.

In Figure 10a, the top-1 and -2 ranked features from SVM, RF, and DNN models are highlighted. For the three models, mapping of important features lineated the same or similar substructures. Figure 10b provides a complementary view of cumulative positive or negative feature contributions. In this case, RF and DNN models predicted a lower probability of activity (p of ∼0.60) than the SVM model (p = 0.97), which largely resulted from negative feature contributions, especially for DNN, which were absent in the SVM model. SHAP results suggest that RF and DNN models made use of the absence of some features to discriminate between active and inactive training compounds. However, such prioritization had a negative impact on the model output for this exemplary active test compound, leading to a lower output probability. Accordingly, a noninformative bias in the training set was likely exploited by these two ML models. For example, both models penalized the absence of feature #12 (SMARTS pattern: [#6D4v4+0H0R], SMILES: C), which was present in 91% of the positive and only in the 8% of the negative training compounds. The representation also shows that the majority of features with positive contributions to the prediction of activity were conserved.

Comparison of Structural Analogs. Analog series provide interesting test cases for local model diagnostics. In most cases, analogs from the same series are predicted to be active because of their high structural similarity. However, there can be exceptions where small structural differences between compounds abruptly change the predicted probability of activity. Such incorrect predictions are of particular interest to better understand intrinsic limitation of activity predictions, provided the underlying models can be interpreted. Figure 11 presents the SHAP analysis of SVM predictions for two histamine H3 receptor antagonists with comparable potency (having pK_i values of 6.2 and 6.3, respectively). One was predicted correctly, the other incorrectly. Figure 11a shows the ECFP4 features with the highest positive and negative contributions on predicted activity. The first analog was accurately predicted (p = 0.98), but the second was not (p =

DOI: 10.1021/acs.jmedchem.9b01101
J. Med. Chem. 2020, 63, 8761−8777



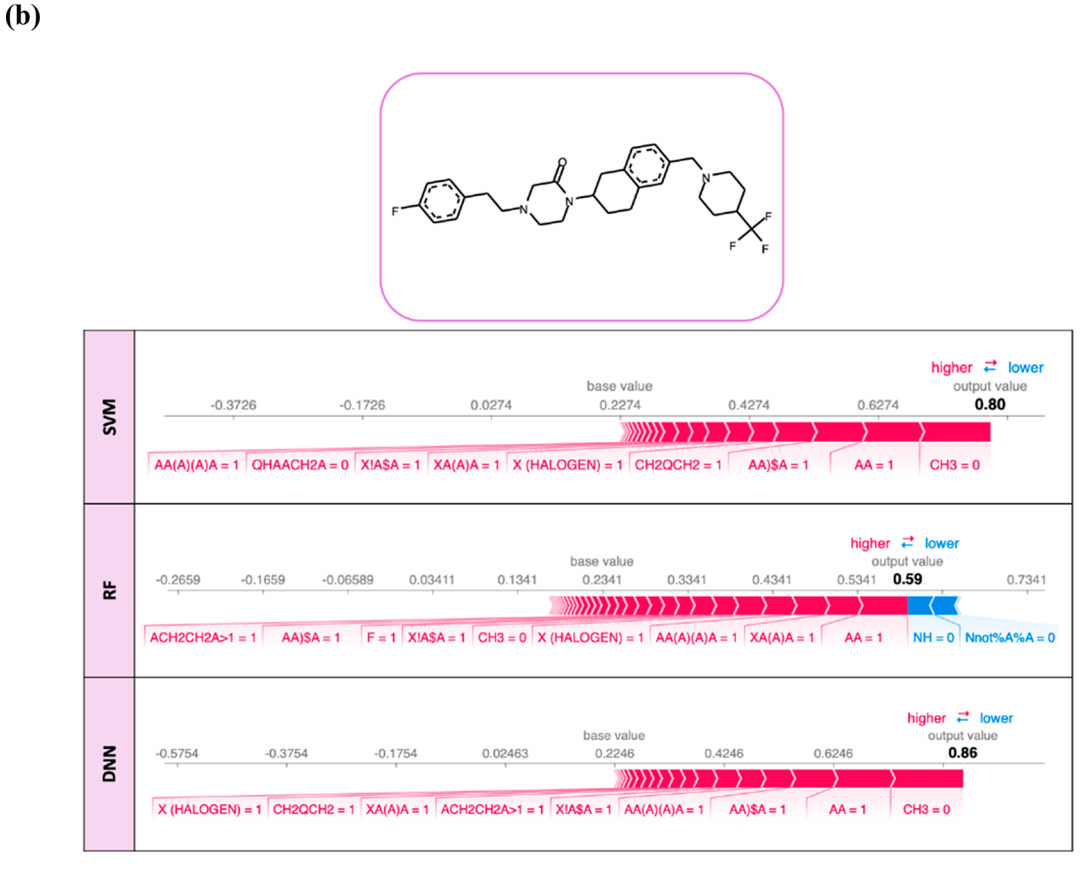


Figure 8. continued

DOI: 10.1021/acs.jmedchem.9b01101
J. Med. Chem. 2020, 63, 8761−8777

(c)

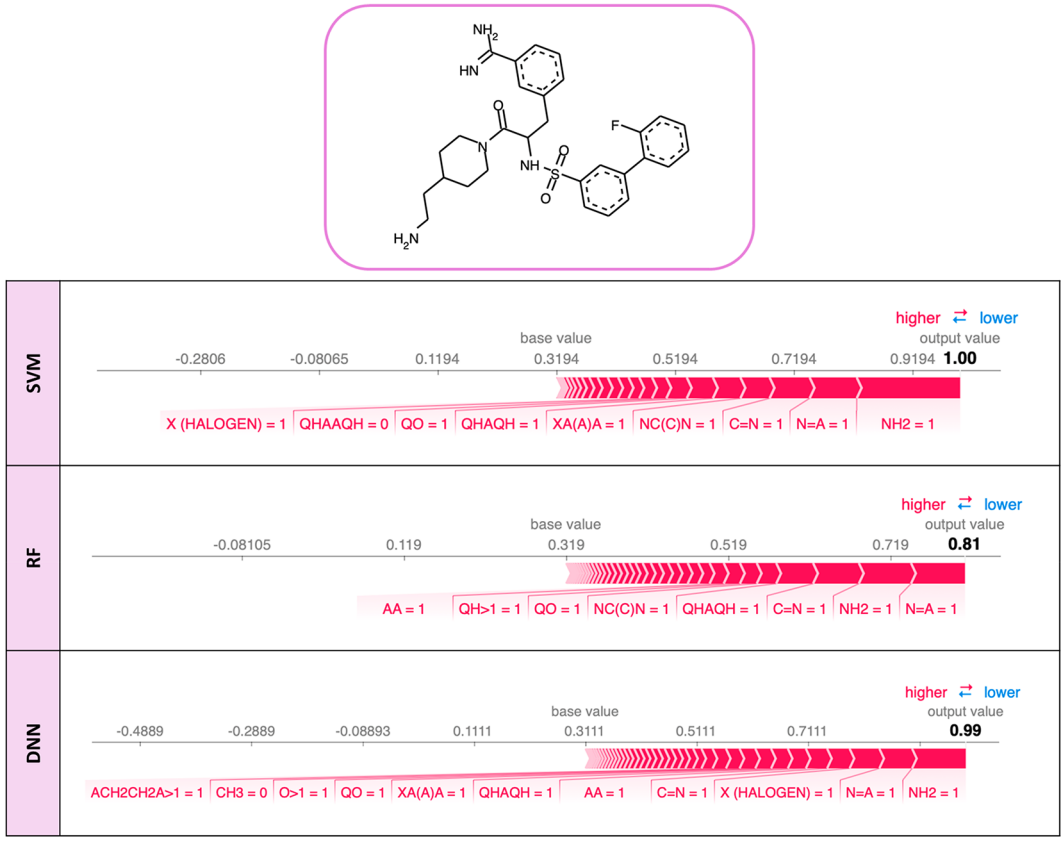


Figure 8. SHAP values for MACCS keys. Shown are two exemplary test compounds that were represented using MACCS keys and correctly predicted by SVM, RF, and DNN models including a (a) Bcl-2 inhibitor, (b) melanin-concentrating hormone receptor 1 antagonist, and (c) factor X inhibitor. SHAP positive (red) and negative (blue) feature weights are given for the three models. The expected base and output value (bold) is also shown. The following symbols are used: A, any element symbol; Q, heteroatom; X, other than H, C, N, O, Si, P, S, F Cl, Br, I; $, ring bond; !, aliphatic bond; %, aromatic bond.

0.19). The substructure formed by features with the highest positive contribution was shared by both compounds but obtained a larger SHAP value for the first analog. Moreover, the correctly predicted compound did not yield any feature with a negative contribution. By contrast, for the incorrectly predicted analog, features making a large negative contribution were identified. Consequently, the substructure formed by features with largest negative contribution according to the SVM model was only present in the incorrectly predicted analog. Figure 11b reports the base and SHAP values for the two compound predictions. Even though most of the variables with positive contributions were shared by both compounds, the second analog exhibited a number of features that negatively impacted the prediction. Thus, SHAP analysis uncovered a model error and made it possible to rationalize why these two analogs produced different model outputs. On the basis of such insights, it can be attempted to further optimize SVM models for individual predictions.

Global Model Diagnostics. SHAP analysis can conveniently be used as a global model diagnostic by comparing decisions of different ML models on the same compound data set, which aids in model selection. Moreover, consensus features can be identified across methodologically distinct models that can be selected for practical applications. Figure 12 presents an example of SHAP-based model comparison and selection. Figure 12a shows a score plot of predicted probabilities of activity for compounds using DNN and SVM models. Red dots in the upper-right panel represent active compounds that are correctly predicted by both methods, and blue dots in the bottom-left panel are inactive compounds correctly detected by SVM and DNN. The compounds falling into other regions of the plot have been incorrectly predicted by only one of the methods. An exemplary active compound that is correctly predicted by DNN but not by SVM is indicated. In Figure 12b, the SHAP contribution plots are shown for this compound and the SVM and DNN models. It is evident that many features were equally weighted using SHAP for predictions with both models. However, SVM was found to assign negative contributions to a number of atom environments that were not considered by DNN. To further reduce the black-box character of these model predictions, highly weighted features were mapped onto this compound, as depicted in Figure 12c. The SHAP threshold was adjusted such that top-1 as well as -3 ranked features with positive contributions were obtained from both SVM and DNN models. For SVM, the top-ranked features with negative contributions were also selected. Such features were absent in the DNN model, as discussed above. Figure 12c shows that features important for the prediction of activity mapped to the same region in the molecule. However, SVM also negatively weighted similar parts of the compound formed by overlapping atom environments, thus reducing the output probability.

DOI: 10.1021/acs.jmedchem.9b01101
J. Med. Chem. 2020, 63, 8761−8777

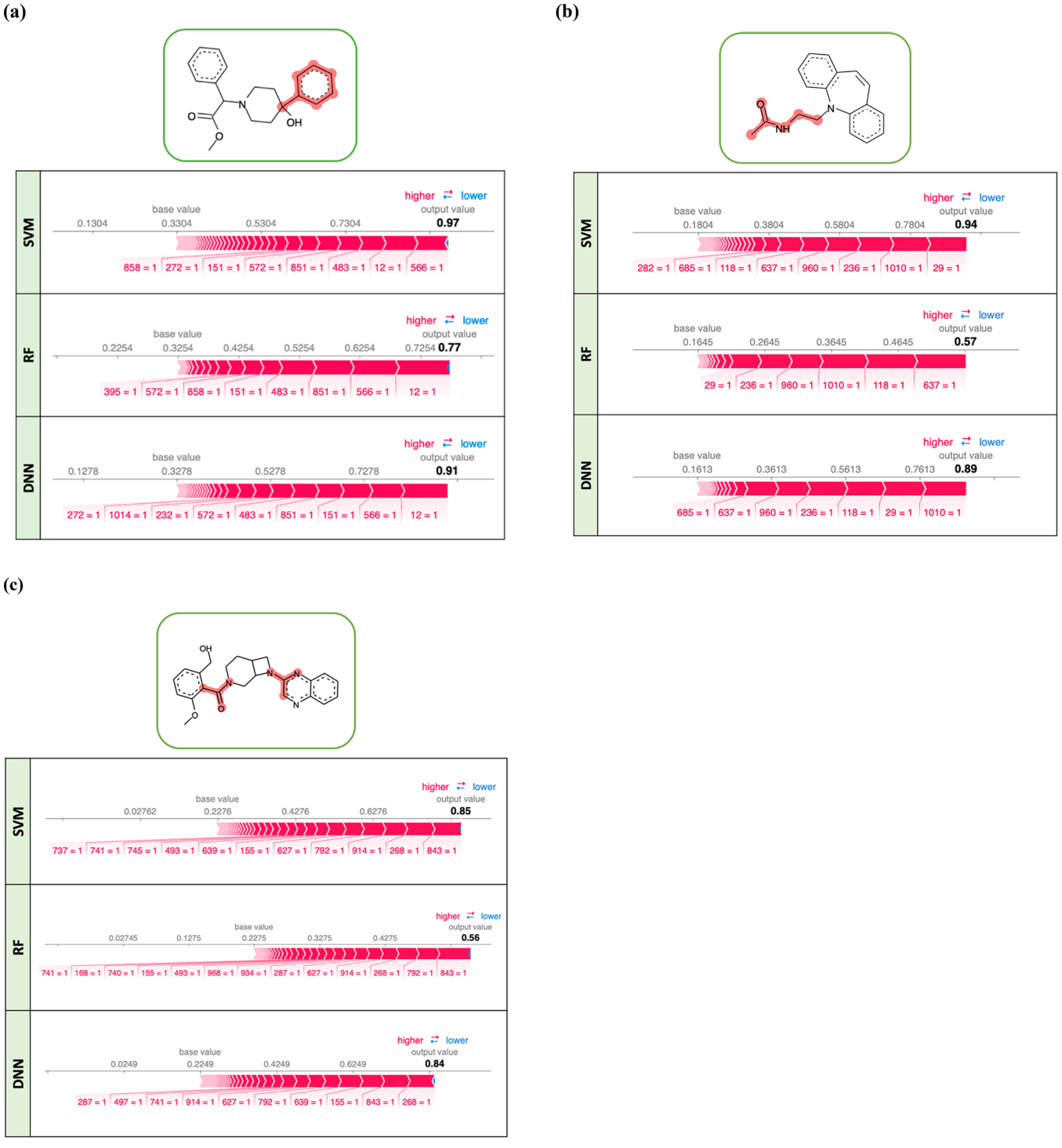


Figure 9. SHAP values for ECFP4 features. Shown are two exemplary test compounds that were represented using ECFP4 and correctly predicted by SVM, RF, and DNN models including a ligand of the (a) κ opioid receptor, (b) melanocortin receptor 1B, and (c) orexin receptor 2. The representation is according to Figure 8. In addition, top-ranked features are highlighted in compound structures.

Thus, in this case, the model diagnostic detected SVM-dependent inconsistencies in feature prioritization, which were absent in the DNN model. On the basis if these observations, the DNN model would be prioritized.

CONCLUSIONS

In this work, the SHAP method has been introduced for the interpretation of compound activity predictions using ML models, regardless of their complexity. As an ML model diagnostic, SHAP is generally applicable to ML models including ensemble and DL models, which makes it possible to shed light on their black-box nature. SHAP values quantify feature importance for ML in a consistent manner. Furthermore, the SHAP analysis scheme introduced herein provides visual access to feature importance and enables structural interpretation of ML predictions including DNNs. By application of the SHAP methodology, variables with increasing influence on predictions can be explored and detect

DOI: 10.1021/acs.jmedchem.9b01101
J. Med. Chem. 2020, 63, 8761−8777

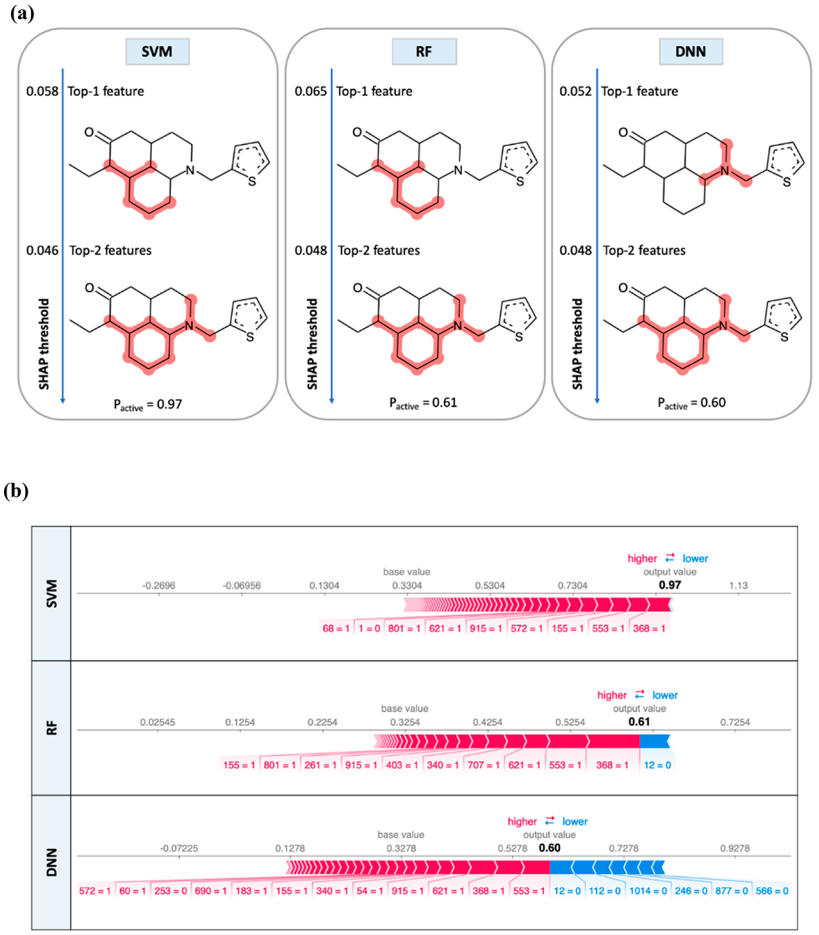


Figure 10. SHAP visualization for ECFP4. SHAP results are shown for an exemplary κ opioid receptor antagonist. In (a), the probability of activity predicted by SVM (left), RF (center), and DNN (right) is reported at the bottom of the boxes and the most important features for determining these predictions (top-1 and top-2) according to SHAP analysis are mapped onto the compound and highlighted. For top-ranked features, the corresponding SHAP values are reported. In (b), positive (red) and negative (blue) feature contributions are shown for SVM (top), RF (middle), and DNN (bottom). The output value (bold) corresponds to the output probability of each ML model.

potential sources of bias of predictions or confirm their consistency and further validate a model. It is important to consider the applicability domain of explanatory methods because interpretations will be strongly influenced by training data and conditions. Here, it is important to note that the SHAP methodology is applicable to essentially all ML approaches including regression techniques. For ML methods and especially in the context of DL, SHAP offers novel opportunities for the rationalization of predictive models and for reducing or eliminating their black-box character. In future work, SHAP analysis might be further extended to better understand multitask learning for compound activity prediction.

EXPERIMENTAL SECTION

1. Compound Data Sets. Machine learning inevitably depends on compounds from the literature and their reported activity data. ChEMBL is the primary repository for active compounds from the medicinal chemistry literature.[44] From ChEMBL version 24, 10 activity classes were selected for ML.

For each selected compound, literature reference(s) and the presence of direct interactions (i.e., assay relationship type "D") with a human single-protein target at the highest confidence level (i.e., assay confidence score 9) were required. As potency measurements, explicitly specified (assay-independent) equilibrium constants (K_i values) were required. Activity measurements provided in ChEMBL were taken from original publications. When multiple K_i values were available for a compound and fell within the same order of magnitude, the mean value was determined. If differences between measurements exceeded 1 order of magnitude, the compound was discarded. Only compounds with (mean) pK_i of at least 5 were ultimately selected to exclude borderline active compounds from further consideration. Furthermore, compounds with potentially inconsistent activity records including comments such as "inactive", "inconclusive", or "not active" were discarded. Taken together, these criteria exclusively select compounds with highest ChEMBL confidence scores and highest activity data confidence.[45] In addition, all compounds meeting high-confidence selection criteria were screened for pan-assay interference compounds (PAINS)[46] using substructure libraries from public filters[44,47,48] and compounds with PAINS alerts were discarded (less than 1%).

Selected data sets were required to contain at least 200 compounds belonging to at least 50 different analog series computationally determined[49] on the basis of matched molecular pair (MMP) relationships.[50] Selection of activity classes consisting of large numbers of analog series ensured the presence of defined subsets of structurally analogous compounds that were distinct from others. Activity classes of sufficient size and intraclass structural diversity were essential for meaningful ML-based activity modeling. Since this study aimed to detect chemical features determining activity predictions, confirmed activity of compounds against a given target based on high-confidence activity data was another key criterion for an activity class. Table 1 specifies selected classes, which consisted of 243–955 compounds and 57–216 analog series, respectively. To prevent

DOI: 10.1021/acs.jmedchem.9b01101
J. Med. Chem. 2020, 63, 8761–8777

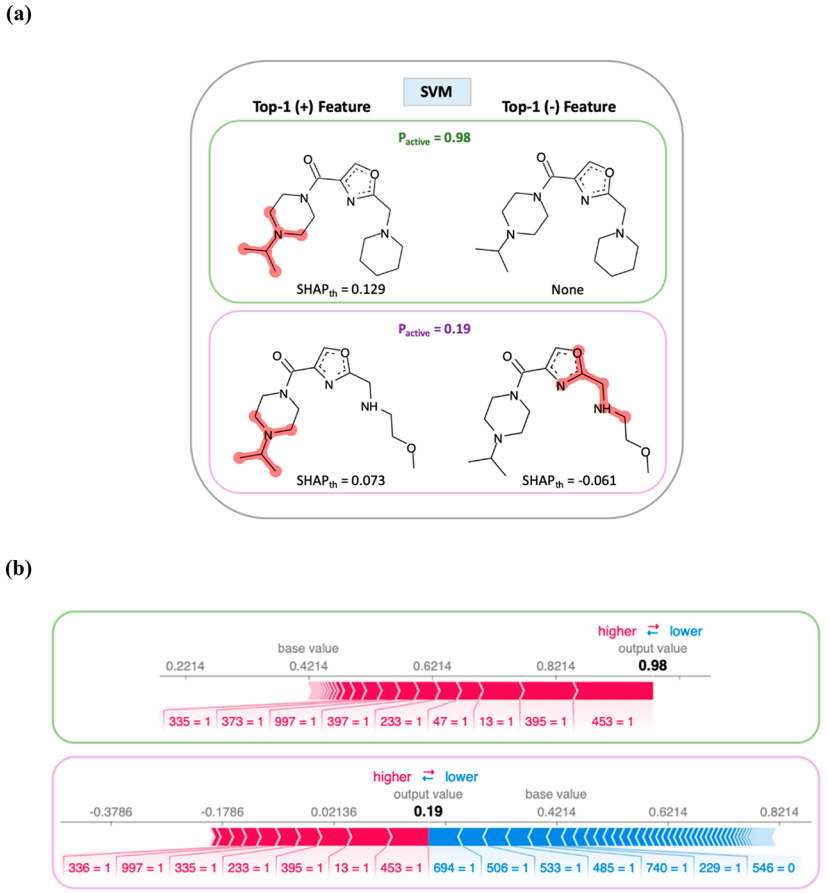


Figure 11. Rationalizing SVM predictions for two analogs. (a) Two analogs are shown (with ECFP4 Tanimoto similarity of 0.6), and features with the largest positive and negative contributions to SVM predictions are highlighted. $SHAP_{th}$ indicates the SHAP threshold value for the top-1 ranked feature (such that only this feature is obtained). The analogs have different predicted probabilities of activity (P_{active}). (b) For the analogs, features with positive (red) and negative (blue) SHAP values are visualized.

potential structural bias of predictions,[51,52] analogs from different series were selected as positive (active) training and test instances. Training sets contained 70% of the analog series per activity class and corresponding test sets 30% of the series. On average, training and test sets included 366 (157 to 683) and 163 (70−278) active compounds, respectively. As negative (inactive) training and test instances, compounds were randomly selected from ZINC,[48] i.e., consistently 1000 compounds per training and test set.

2. Molecular Representations. Extended connectivity fingerprint with bond diameter 4 (ECFP4)[53] is a topology descriptor encoding layered atom environments as numeric identifiers using a hashing function. SMARTS patterns corresponding to each atom environment (codified by a hash value) were stored. Therefore, ECFP4 features can be mapped back onto the compounds. This feature set fingerprint is variable in size, but a constant-length 1024-bit representation was obtained through modulo mapping. In addition, MACCS structural keys[54] were used in a binary fingerprint format encoding the presence (bit set on) or absence (off) of 166 predefined structural patterns or fragments. The OEChem toolkit[55] and in-house Python scripts were used for fingerprint calculations.

3. Machine Learning Models. *3.1. Support Vector Machine.* The SVM classifier finds a hyperplane in a multidimensional space that maximizes the distance between the support vectors of each class, known as *margin*.[17] The support vectors are the training instances of one class that are closest to the other class. SVM enables nonlinear modeling through the application of the *kernel trick*,[56] i.e., the use of kernel functions to map training compounds into a higher-dimensional feature space representation in which the classes might be linearly separable. For compound classification, the nonlinear Tanimoto kernel[54] is one of the best performing kernel functions.[57,58] The SVM implementation of scikit-learn[56] with customized Tanimoto kernel was used for all calculations.

3.2. Random Forest. RF is an ensemble of decision trees (DTs) that aims at reducing the variance of individual trees.[16] RF is based on bootstrap aggregating according to which training DTs with distinct compound subsets are generated. In addition, a random subset of features is used to minimize correlations between DTs. The final RF prediction results from a consensus across the DT population. RF calculations were carried out with scikit-learn.[59]

3.3. Feedforward Deep Neural Networks. A DNN is a series of functional transformations (neurons) that learn how to modify input values to obtain a desired output.[60] Accordingly, DNNs have an input layer, multiple hidden layers, and an output layer. First, a neuron's input values (x_1, ..., x_D) are linearly combined considering a set of weights ($\mathbf{w}$) and biases (b). Then, a differentiable nonlinear activation function (h) is applied to obtain the neuron's output (y_j) according to eq 4:[61]

$$y_j = h\left(\sum_{i=1}^{D} \omega_{ji}^{(n)} x_i + b_j^{(n)}\right) \tag{4}$$

where n indicates the layer number. Training aims at determining the weights and biases that minimize the cost function (e.g., cross-entropy).[21] Gradient descent is applied to update weights by considering small steps (defined by the learning rate) in the direction of the negative gradient and can be efficiently calculated using backpropagation.[60] DNNs were generated using TensorFlow[61] and Keras.[62]

3.4. Hyperparameter Optimization. Model hyperparameters were optimized through internal 2-fold cross-validation and grid search.

DOI: 10.1021/acs.jmedchem.9b01101
J. Med. Chem. 2020, 63, 8761−8777

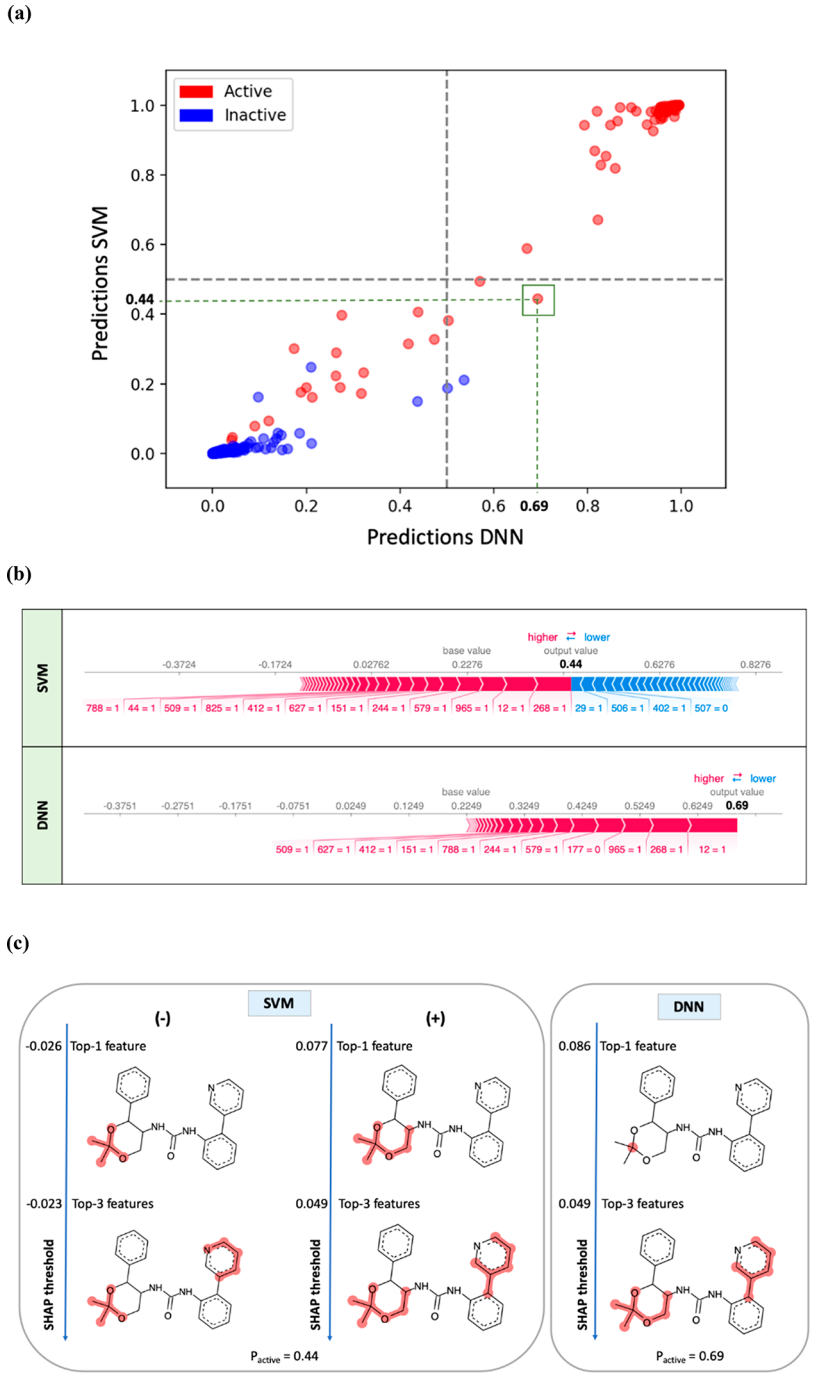


Figure 12. Interpretation of DNN and SVM predictions. (a) Score plots show the output probabilities of activity against orexin receptor 2 for DNN and SVM models. A green square marks an exemplary compound that is incorrectly classified by SVM (p = 0.44) but correctly predicted by DNN (p = 0.69). (b) Plots for SVM and DNN report SHAP feature values that modify the base value (0.22), with a positive (red) or negative (blue) sign, to yield the final output probability (bold). For the DNN model, features with negative contributions to the output probability were absent. (c) ECFP4 features with largest positive and negative SHAP values are shown for the SVM model (i.e., the top-1 feature and the top-3 ranked features). For the DNN model, only the features with positive SHAP values are available.

The same randomized data splits were considered for training (80%) and internal validation (20%) for different ML methods.[63] Best hyperparameters were selected according to area under the ROC curve (AUC) optimization (average across folds).

For SVM, the regularization term C was optimized with candidate values of 0.01, 0.1, 1, and 10. In addition, SVM models were built with and without class weights.[58] The use of class weights consists in penalizing errors on the minority class more than errors on the majority class.

For RF models, the number of trees was consistently set to 500 and three numerical hyperparameters were optimized including the minimum number of samples required to split a leaf node (1, 5, 10) or an internal node (2, 8, 16) and the maximum number of features considered when searching for the best split (i.e., square root, $\log_2$). Furthermore, models were built with and without class weights.

Different network architectures were tested for DNN models, with the following number of neurons in hidden layers: [100,500], [200,100], [2000,1000], [200,100,100], and [2000,1000,100]. The activation function was Rectified Linear Unit (ReLU) except at the

DOI: 10.1021/acs.jmedchem.9b01101
J. Med. Chem. 2020, 63, 8761−8777

output layer, where a sigmoid function was applied. In addition, three initial learning rates (0.1, 0.01, 0.001) were tested and values were reduced when reaching a loss plateau. L2 regularization and drop-out (25% or 50%) were applied to all hidden layers. Three batch sizes (64, 128, 256) were tested, Adam was used as the optimization function, and the number of epochs was set to 50 and 200 during internal and external validation, respectively.

3.5. Performance Measures. Predictive performance on test sets was evaluated using three metrics: AUC, balanced accuracy (BA),[64] and Matthew's correlation coefficient (MCC).[65] BA and MCC are defined by eqs 5 and 6, respectively.

$$\mathrm{BA} = \frac{1}{2}(\mathrm{TPR} + \mathrm{TNR}) \tag{5}$$

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}} \tag{6}$$

To statistically compare MCC values before and after feature elimination, nonparametric Wilcoxon tests[66] were carried out.

4. Feature Contributions. Feature contributions were assessed following the SHAP approach detailed in the Results sections. The feature contributions represented by Shapley values are meant to satisfy three axioms including *local accuracy*, *consistency*, and *nonexistence* (or null effect).[67,68]

5. Data Availability. Compound activity classes used here are made available in an open access deposition on the ZENODO platform.[69]

AUTHOR INFORMATION

Corresponding Author

*Phone: +49-228-7369-100. Fax: +49-228-7369-101. E-mail: bajorath@bit.uni-bonn.de.

ORCID

Jürgen Bajorath: 0000-0002-0557-5714

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The project leading to this report has received funding (for R.R.-P.) from the European Union's Horizon 2020 Research and Innovation program under the Marie Skłodowska-Curie Grant Agreement 676434, "Big Data in Chemistry" ("BIG-CHEM", http://bigchem.eu). The article reflects only the authors' view, and neither the European Commission nor the Research Executive Agency (REA) is responsible for any use that may be made of the information it contains. The authors thank OpenEye Scientific Software, Inc., for providing a free academic license for the OpenEye toolkit, and Scott Lundberg for the SHAP library. The authors also thank Dagmar Stumpfe and Martin Vogt for help with compound data analysis.

ABBREVIATIONS USED

AUC, area under the ROC curve; BA, balanced accuracy; DL, deep learning; DNN, deep neural network; DT, decision tree; ECFP, extended connectivity fingerprint; LIME, local interpretable model-agnostic explanations; MCC, Matthew's correlation coefficient; ML, machine learning; PAINS, pan-assay interference compounds; (Q)SAR, (quantitative) structure–activity relationship; RF, random forest; SHAP, Shapley additive explanations; SVM, support vector machine

REFERENCES

(1) Varnek, A.; Baskin, I. Machine Learning Methods for Property Prediction in Cheminformatics: Quo Vadis? *J. Chem. Inf. Model.* **2012**, *52*, 1413−1437.

(2) Rodríguez-Pérez, R.; Miyao, T.; Jasial, S.; Vogt, M.; Bajorath, J. Prediction of Compound Profiling Matrices Using Machine Learning. *ACS Omega* **2018**, *3*, 4713−4723.

(3) Lavecchia, A. Machine-Learning Approaches in Drug Discovery: Methods and Applications. *Drug Discovery Today* **2015**, *20*, 318−331.

(4) Lo, Y.; Rensi, S. E.; Torng, W.; Altman, R. B. Machine Learning in Chemoinformatics and Drug Discovery. *Drug Discovery Today* **2018**, *23*, 1538−1546.

(5) Cherkasov, A.; Muratov, E.; Fourches, D.; Varnek, A.; Baskin, I. I.; Cronin, M.; Dearden, J.; Gramatica, P.; Martin, Y. C.; Todeschini, R.; Consonni, V.; Kuz'min, V. E.; Cramer, R.; Benigni, R.; Yang, C.; Rathman, J.; Terfloth, L.; Gasteiger, J.; Richard, A.; Tropsha, A. QSAR Modeling: Where Have You Been? Where Are You Going to? *J. Med. Chem.* **2014**, *57*, 4977−5010.

(6) Zakharov, A. V.; Peach, M. L.; Sitzmann, M.; Nicklaus, M. C. QSAR Modeling of Imbalanced High-Throughput Screening Data in PubChem. *J. Chem. Inf. Model.* **2014**, *54*, 705−712.

(7) Lewis, R. A. A General Method for Exploiting QSAR Models in Lead Optimization. *J. Med. Chem.* **2005**, *48*, 1638−1648.

(8) Bajorath, J. Integration of Virtual and High-throughput Screening. *Nat. Rev. Drug Discovery* **2002**, *1*, 882−894.

(9) Klon, A. E.; Lowrie, J. F.; Diller, D. J. Improved Naïve Bayesian Modeling of Numerical Data for Absorption, Distribution, Metabolism and Excretion (ADME) Property Prediction. *J. Chem. Inf. Model.* **2006**, *46*, 1945−1956.

(10) Guha, R. On the Interpretation and Interpretability of Quantitative Structure-Activity Relationship Models. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 857−871.

(11) Doweyko, A. M. QSAR: Dead or Alive? *J. Comput.-Aided Mol. Des.* **2008**, *22*, 81−89.

(12) Sieg, J.; Flachsenberg, F.; Rarey, M. In the Need of Bias Control: Evaluating Chemical Data for Machine Learning in Structure-Based Virtual Screening. *J. Chem. Inf. Model.* **2019**, *59*, 947−961.

(13) Polishchuk, P. Interpretation of Quantitative Structure-Activity Relationship Models: Past, Present, and Future. *J. Chem. Inf. Model.* **2017**, *57*, 2618−2639.

(14) Hansen, K.; Baehrens, D.; Schroeter, T.; Rupp, M.; Müller, K.-R. Visual Interpretation of Kernel-Based Prediction Models. *Mol. Inf.* **2011**, *30*, 817−826.

(15) Balfer, J.; Bajorath, J. Introduction of a Methodology for Visualization and Graphical Interpretation of Bayesian Classification Models. *J. Chem. Inf. Model.* **2014**, *54*, 2451−2468.

(16) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5−32.

(17) Vapnik, V. N. *The Nature of Statistical Learning Theory*, 2nd ed.; Springer: New York, 2000.

(18) Balfer, J.; Bajorath, J. Visualization and Interpretation of Support Vector Machine Activity Predictions. *J. Chem. Inf. Model.* **2015**, *55*, 1136−1147.

(19) Hu, Y.; Bajorath, J. Learning from 'Big Data': Compounds and Targets. *Drug Discovery Today* **2014**, *19*, 357−360.

(20) Papadatos, G.; Gaulton, A.; Hersey, A.; Overington, J. P. Activity, Assay and Target Data Curation and Quality in the ChEMBL database. *J. Comput.-Aided Mol. Des.* **2015**, *29*, 885−896.

(21) Nielsen, M. A. *Neural Networks and Deep Learning*; Determination Press, 2015.

(22) Krizhevsky, A.; Sutskever, I.; Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. *Adv. Neural Inf. Process. Syst.* **2015**, *25*, 1097−1105.

(23) Hinton, G. E.; Deng, L.; Yu, D.; Dahl, G.; Mohamed, A.; Jaitly, N.; Senior, A.; Vanhoucke, V.; Nguyen, P.; Sainath, T.; Kingsbury, B. Deep Neural Networks for Acoustig Modeling in Speech Recognition. *IEEE Signal Process. Mag.* **2012**, *29*, 82−97.

DOI: 10.1021/acs.jmedchem.9b01101
J. Med. Chem. 2020, 63, 8761−8777

(24) Baskin, I.; Winkler, D.; Tetko, I. V. A Renaissance of Neural Networks in Drug Discovery. *Expert Opin. Drug Discovery* **2016**, *11*, 785−795.
(25) Ekins, S. The Next Era: Deep Learning in Pharmaceutical Research. *Pharm. Res.* **2016**, *33*, 2594−2603.
(26) Ma, J.; Sheridan, R. P.; Liaw, A.; Dahl, G. E.; Svetnik, V. Deep Neural Nets as a Method for Quantitative Structure-Activity Relationships. *J. Chem. Inf. Model.* **2015**, *55*, 263−274.
(27) Lenselink, E. B.; ten Dijke, N.; Bongers, B.; Papadatos, G.; van Vlijmen, H. W. T.; Kowalczyk, W.; Ijzerman, A. P.; van Westen, G. J. P. Beyond the Hype: Deep Neural Networks Outperform Established Methods Using a ChEMBL Bioactivity Benchmark Set. *J. Cheminf.* **2017**, *9*, No. e45.
(28) Mayr, A.; Klambauer, G.; Unterthiner, T.; Steijaert, M.; Wegner, J. K.; Ceulemans, H.; Clevert, D.; Hochreiter, S. Large-Scale Comparison of Machine Learning Methods for Drug Target Prediction on ChEMBL. *Chem. Sci.* **2018**, *9*, 5441−5451.
(29) Ramsundar, B.; Liu, B.; Wu, Z.; Verras, A.; Tudor, M.; Sheridan, R. P.; Pande, V. Is Multitask Deep Learning Practical for Pharma? *J. Chem. Inf. Model.* **2017**, *57*, 2068−2076.
(30) Xu, Y.; Ma, J.; Liaw, A.; Sheridan, R. P.; Svetnik, V. Demystifying Multitask Deep Neural Networks for Quantitative Structure-Activity Relationships. *J. Chem. Inf. Model.* **2017**, *57*, 2490−2504.
(31) Rodríguez-Pérez, R.; Bajorath, J. Prediction of Compound Profiling Matrices, Part II: Relative Performance of Multitask Deep Learning and Random Forest Classification on the Basis of Varying Amounts of Training Data. *ACS Omega* **2018**, *3*, 12033−12040.
(32) Lundberg, S. M.; Nair, B.; Vavilala, M. S.; Horibe, M.; Eisses, M. J.; Adams, T.; Liston, D. E.; Low, D. K.; Newman, S.; Kim, J.; Lee, S. Explainable Machine-learning Predictions for the Prevention of Hypoxaemia During Surgery. *Nat. Biomed. Eng.* **2018**, *2*, 749−760.
(33) Rodríguez-Pérez, R.; Vogt, M.; Bajorath, J. Support Vector Machine Classification and Regression Prioritize Different Structural Features for Binary Compound Activity and Potency Value Prediction. *ACS Omega* **2017**, *2*, 6371−6379.
(34) Hastie, T.; Tibshirani, R.; Friedman, J. H. *The Elements of Statistical Learning*; Springer: Berlin, 2009.
(35) Iooss, B.; Saltelli, A. Introduction to Sensitivity Analysis. In *Handbook of Uncertainty Quantification*; Ghanem, R., Higdon, D., Owhadi, H., Eds.; Springer International Publishing: Cham, Switzerland, 2016; pp 1−20.
(36) So, S. S.; Richards, W. G. Application of Neural Networks: Quantitative Structure-Activity Relationships of the Derivatives of 2,4-Diamino-5-(Substituted-Benzyl)Pyrimidines as DHFR Inhibitors. *J. Med. Chem.* **1992**, *35*, 3201−3207.
(37) Baskin, I. I.; Ait, A. O.; Halberstam, N. M.; Palyulin, V. A.; Zefirov, N. S. An Approach to the Interpretation of Backpropagation Neural Network Models in QSAR Studies. *SAR QSAR Environ. Res.* **2002**, *13*, 35−41.
(38) Marcou, G.; Horvath, D.; Solov'ev, V.; Arrault, A.; Vayer, P.; Varnek, A. Interpretability of SAR/QSAR Models of Any Complexity by Atomic Contributions. *Mol. Inf.* **2012**, *31*, 639−642.
(39) Fujita, T.; Winkler, D. A. Understanding the Roles of the "Two QSARs". *J. Chem. Inf. Model.* **2016**, *56*, 269−274.
(40) Johansson, U.; Sönströd, C.; Norinder, U.; Boström, H. Trade-Off between Accuracy and Interpretability for Predictive in Silico Modeling. *Future Med. Chem.* **2011**, *3*, 647−663.
(41) Lundberg, S. M.; Lee, S. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30*; NIPS, 2017.
(42) Ribeiro, M. T.; Singh, S.; Guestrin, C. "Why Should I Trust You?"Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* **2016**, 1135−1144.
(43) Shapley, L. S. A Value for N-Person Games. In *Contributions to the Theory of Games*; Kuhn, H. W., Tucker, A. W., Eds.; Annals of Mathematical Studies; Princeton University Press, 1953; pp 307−317.
(44) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: A Large-scale Bioactivity Database for Drug Discovery. *Nucleic Acids Res.* **2012**, *40*, D1100−D1107.
(45) Hu, Y.; Bajorath, J. Influence of Search Parameters and Criteria on Compound Selection, Promiscuity, and Pan Assay Interference Characteristics. *J. Chem. Inf. Model.* **2014**, *54*, 3056−3066.
(46) Baell, J. B.; Holloway, G. A. New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for Their Exclusion in Bioassays. *J. Med. Chem.* **2010**, *53*, 2719−2740.
(47) RDKit: Cheminformatics and Machine Learning Software. 2013. http://www.rdkit.org (accessed June 3, 2019).
(48) Sterling, T.; Irwin, J. J. ZINC 15–Ligand Discovery for Everyone. *J. Chem. Inf. Model.* **2015**, *55*, 2324−2337.
(49) Stumpfe, D.; Dimova, D.; Bajorath, J. Computational Method for the Systematic Identification of Analog Series and Key Compounds Representing Series and Their Biological Activity Profiles. *J. Med. Chem.* **2016**, *59*, 7667−7676.
(50) Kenny, P. W.; Sadowski, J. Structure Modification in Chemical Databases. In *Cheminformatics in Drug Discovery*; Oprea, T. I., Ed.; Wiley-VCH: Weinheim, Germany, 2004; pp 271−285.
(51) Scior, T.; Bender, A.; Tresadern, G.; Medina-Franco, J. L.; Martínez-Mayorga, K.; Langer, T.; Cuanalo-Contreras, K.; Agrafiotis, D. K. Recognizing Pitfalls in Virtual Screening: A Critical Review. *J. Chem. Inf. Model.* **2012**, *52*, 867−881.
(52) Rodríguez-Pérez, R.; Bajorath, J. Multitask Machine Learning for Classifying Highly and Weakly Potent Kinase Inhibitors. *ACS Omega* **2019**, *4*, 4367−4375.
(53) Rogers, D.; Hahn, M. Extended Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.
(54) *MACCS Structural Keys*; Accelrys: San Diego, CA, 2011.
(55) *OEChem TK*, version 2.0.0; OpenEye Scientific Software: Santa Fe, NM, 2015.
(56) Boser, B. E.; Guyon, I. M.; Vapnik, V. N. A Training Algorithm for Optimal Margin Classifiers. *Proceedings of the 5th Annual Workshop on Computational Learning Theory; Pittsburgh, Pennsylvania, 1992*; ACM: New York, 1992; pp 144−152.
(57) Ralaivola, L.; Swamidass, S. J.; Saigo, H.; Baldi, P. Graph Kernels for Chemical Informatics. *Neural Netw* **2005**, *18*, 1093−1110.
(58) Rodríguez-Pérez, R.; Vogt, M.; Bajorath, J. Influence of Varying Training Set Composition and Size on Support Vector Machine-Based Prediction of Active Compounds. *J. Chem. Inf. Model.* **2017**, *57*, 710−716.
(59) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.
(60) Bishop, C. M. *Pattern Recognition and Machine Learning*; Springer: New York, 2006.
(61) Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M.; Kudlur, M.; Levenberg, J.; Monga, R.; Moore, S.; Murray, D. G.; Steiner, B.; Tucker, P.; Vasudevan, V.; Warden, P.; Wicke, M., Yu, Y.; Zheng, X. TensorFlow: A System for Large-scale Machine Learning. Presented at the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16), Savannah, GA, 2016.
(62) Chollet, F. Keras, version 2.1.3, 2015. https://github.com/keras-team/keras (accessed January 17, 2018).
(63) Baumann, D.; Baumann, K. Reliable Estimation of Prediction Errors for QSAR Models under Model Uncertainty Using Double Cross-Validation. *J. Cheminf.* **2014**, *6*, No. e47.
(64) Brodersen, K. H.; Ong, C. S.; Stephan, K. E.; Buhmann, J. M. The Balanced Accuracy and Its Posterior Distribution. *Proceedings of the 20th International Conference on Pattern Recognition (ICPR)* **2010**, 3121−3124.

DOI: 10.1021/acs.jmedchem.9b01101
J. Med. Chem. 2020, 63, 8761−8777

(65) Matthews, B. Comparison of the Predicted and Observed Secondary Structure of T4 Phage Lysozyme. *Biochim. Biophys. Acta, Protein Struct.* **1975**, *405*, 442−451.
(66) Conover, W. J. On Methods of Handling Ties in the Wilcoxon Signed-Rank Test. *J. Am. Stat. Assoc.* **1973**, *68*, 985−988.
(67) Osborne, M. J.; Rubinstein, A. *A Course in Game Theory*; The MIT Press: Cambridge, MA, 1994.
(68) Young, H. P. Monotonic Solutions of Cooperative Games. *Int. J. Game Theory.* **1985**, *14*, 65−72.
(69) Rodríguez-Pérez, R.; Bajorath, J. Compound Activity Classes from ChEMBL for Machine Learning Analysis. https://zenodo.org/record/3362353#.XUrdh5MzafU.

DOI: 10.1021/acs.jmedchem.9b01101
J. Med. Chem. 2020, 63, 8761−8777