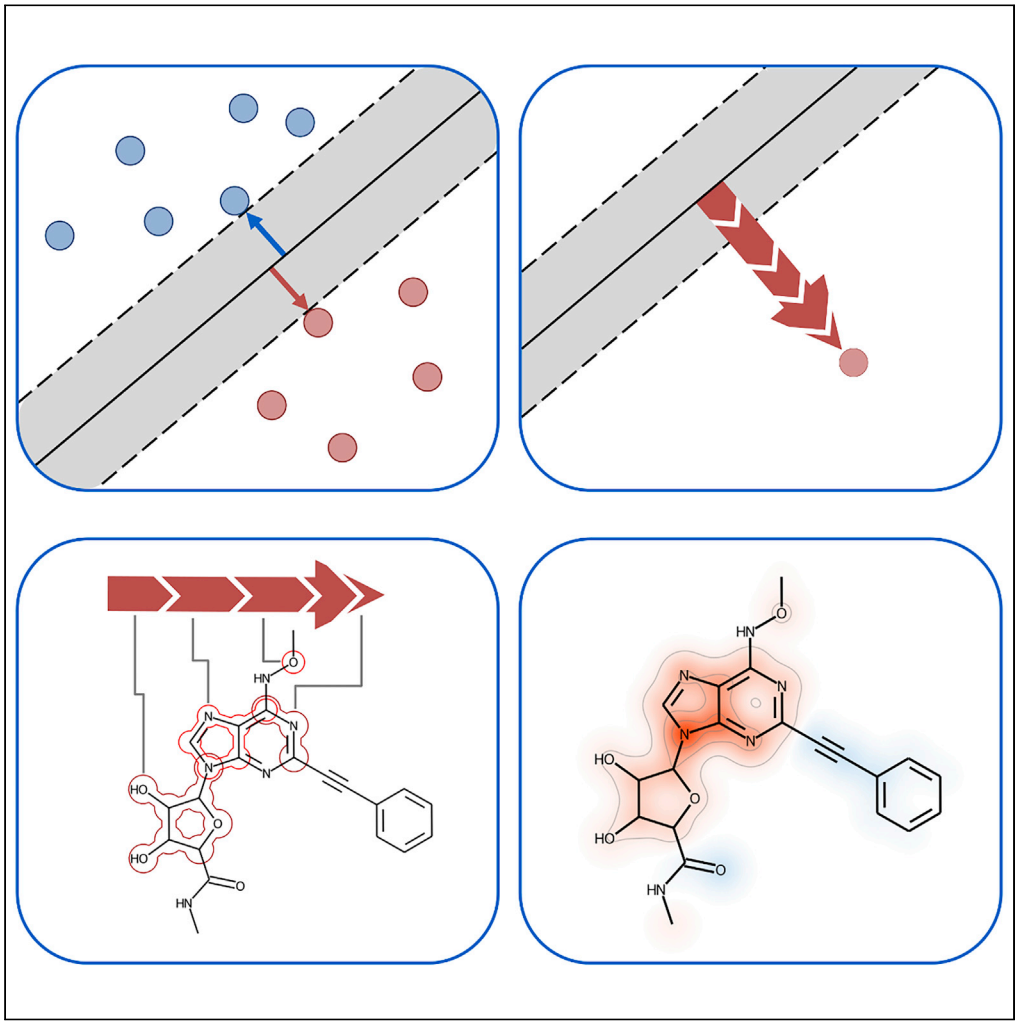
iScience

CellPress
OPEN ACCESS

Article

Calculation of exact Shapley values for support vector machines with Tanimoto kernel enables model interpretation



Christian Feldmann, Jürgen Bajorath

bajorath@bit.uni-bonn.de

Highlights

SVETA: new methodology for explaining support vector machine (SVM) predictions

Tanimoto similarity-based SVM models are popular in chemistry

SVETA enables the calculation of exact Shapley values for rationalizing SVM models

SVETA-based feature mapping provides intuitive explanations of SVM decisions

Feldmann & Bajorath, iScience *25*, 105023
September 16, 2022
https://doi.org/10.1016/j.isci.2022.105023

iScience

CellPress
OPEN ACCESS

Article

Calculation of exact Shapley values for support vector machines with Tanimoto kernel enables model interpretation

Christian Feldmann[1] and Jürgen Bajorath[1,2,*]

SUMMARY

The support vector machine (SVM) algorithm is popular in chemistry and drug discovery. SVM models have black box character. Their predictions can be interpreted through feature weighting or the model-agnostic Shapley additive explanations (SHAP) formalism that locally approximates Shapley values (SVs) originating from game theory. We introduce an algorithm termed SV-expressed Tanimoto similarity (SVETA) for the exact calculation of SVs to explain SVM models employing the Tanimoto kernel, the gold standard for the assessment of molecular similarity. For a model system, the exact calculation of SVs is demonstrated. In an SVM-based compound classification task from drug discovery, only a limited correlation between exact SV and SHAP values is observed, prohibiting the use of approximate values for rationalizing predictions. For exemplary test compounds, atom-based mapping of prioritized features delineates coherent substructures that closely resemble those obtained by analyzing independently derived random forest models, thus providing consistent explanations.

INTRODUCTION

Machine learning (ML) methods are an important component of chemoinformatics and computer-aided drug discovery, especially as the volumes of data available for learning continue to grow (Chen et al., 2018; Varnek and Baskin, 2012). ML infers drug discovery-relevant properties of novel compounds by the statistical assessment of structure-property patterns derived from known molecules (Lo et al., 2018) and contributes to the prioritization of promising candidates (Lavecchia, 2015). For predictive modeling, key properties include biological activities (Lo et al., 2018), pharmacokinetic and -dynamic characteristics (Yamashita and Hashida, 2004), or physiochemical properties (Sellwood et al., 2018).

Most non-linear ML methods have a black box character (Castelvecchi, 2016) meaning that their predictions cannot be intuitively accessed and understood by humans. Although this is unsatisfactory from an intellectual perspective, the black box of ML also limits the acceptance of predictions for experimental design, which represents a serious issue in interdisciplinary research and drug discovery. Therefore, while achieving accurate predictions continues to be the central challenge of ML, increasing emphasis is also put on methodologies for explaining ML models and rationalizing predictions (Vamathevan et al., 2019), often referred to as explainable ML (XML) (Belle and Papantonis, 2021). In most instances, additional algorithms are employed to elucidate decisions of ML models (Štrumbelj and Kononenko, 2014), which can be subdivided into model-specific and model-agnostic approaches (Belle and Papantonis, 2021). Model explanations assist in communicating key findings to non-experts, provide opportunities to gain knowledge of critically important features, and increase confidence in predictions, especially when learned features responsible for model decisions meet chemical intuition (Rodríguez-Pérez and Bajorath, 2021).

Model-agnostic XML approaches are especially thought after as they enable immediate comparisons of different ML models and are also applicable to complex deep learning architectures (LeCun et al., 2015). The Shapley value (SV) formalism from game theory (Shapley, 2016) is an increasingly popular model-agnostic concept for quantifying feature importance for ML predictions (Rodríguez-Pérez and Bajorath, 2021). As originally conceived, SVs provide unique solutions accounting for the contributions of individual players to the performance of a team (Shapley, 2016). In ML settings, players correspond to (representation) features and team performance corresponds to a prediction. Here, SVs are calculated to

[1]Department of Life Science Informatics and Data Science, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Friedrich-Hirzebruch-Allee 5/6, 53115 Bonn, Germany

[2]Lead contact

*Correspondence: bajorath@bit.uni-bonn.de

https://doi.org/10.1016/j.isci.2022.105023

quantify the contributions of features that are present or absent in test instances. The sum of all positive and negative contributions gives the probability of a prediction. As SV calculations depend on the ordering of players/features and are thus combinatorial in nature, the computational costs involved exponentially increase with the number of features. Accordingly, the calculations become essentially infeasible for high-dimensional feature sets typically used in ML. Therefore, for ML, a modification of the SV approach has been introduced termed "Shapley additive explanations" (SHAP) that approximates the exhaustive calculation of SVs in high-dimensional feature spaces by deriving a simplified model in the vicinity of a test instance (Lundberg and Lee, 2017). For this purpose, SHAP makes use of kernel functions and the resulting kernel explainer can be perceived as an extension of the "locally interpretable model-agnostic explanations" (LIME) methodology (Ribeiro et al., 2016). Although the SHAP approximation is generally applicable in ML, another algorithmic variant has been introduced specifically for decision tree-based methods such as random forests (RFs), enabling the calculation of exact SV values (Lundberg et al., 2020). Apart from emerging deep neural networks, decision tree methods and support vector machines (SVMs) are the most widely used ML approaches in bio- and chemoinformatics, drug discovery, and beyond.

In this work, we further extend the SV formalism for ML with another algorithm specifically designed to obtain exact SV values for feature importance in SVM modeling (Cortes and Vapnik, 1995) with the Tanimoto kernel (Ralaivola et al., 2005). In molecular ML, SVMs with Tanimoto kernel are preferentially used (Heikamp and Bajorath, 2014), given the central relevance of the Tanimoto coefficient (Tanimoto, 1958) for the assessment of molecular similarity. The SVM/Tanimoto kernel architecture is almost exclusively employed with binary structural fingerprints as a molecular representation (Willett, 2014), one of the most popular approaches for activity-based compound classification.

SVM solves a binary classification task by constructing a hyperplane in feature space that best separates positive and negative training instances. By maximizing the distance between the hyperplane and learned instances, the generalizability of an SVM model increases and enables the derivation of the separating hyperplane by a limited number of closets instances termed support vectors (Cortes and Vapnik, 1995). If a linearly separating hyperplane cannot be constructed in a given feature space, SVMs project the data into a higher-dimensional representation through the use of kernel functions where linear separation might become feasible. This so-called "kernel-trick" (Boser et al., 1992) does not require explicit mapping of objects from one feature space into another and is thus computationally efficient. Given its characteristic operations, the SVM algorithm, which can be readily adapted for regression tasks (support vector regression; SVR), has black box character. For analyzing predictions using kernel-based ML methods including SVM and SVR, feature or variable weighting and visualization techniques have been developed previously (Balfer and Bajorath, 2015; Sun et al., 2017; Üstün et al., 2007), which are only applicable to features that are present in test instances.

Herein, we introduce the concept of SV-expressed Tanimoto similarity (SVETA) that enables the exact calculation of feature importance values for SVMs with the Tanimoto kernel and the rationalization of their predictions. In ML, the TreeExplainer approach (Lundberg et al., 2020) referred to above and SVETA currently are the only methods available to calculate exact SVs for rationalizing predictions of RF and SVM models, respectively.

RESULTS

Conceptual framework

Following the SV concept, the importance (φ) of a feature (f) is calculated by assessing the difference in model output (v) for input data with or without the feature. This change is determined systematically for each possible subset S (coalition) of remaining features (F\{f}), weighted by the inverse multinomial coefficient. This coefficient is calculated as the number of permutations of the coalition multiplied by the number of permutations of features not contained in the coalition and divided by the total number of feature permutations (F), as defined by Equation 1:

$$\varphi_f(v) = \sum_{S \subseteq F \{f\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} (v(S \cup \{f\}) - v(S)) \quad \text{(Equation 1)}$$

As the model output must be evaluated for each possible coalition computational requirements scale exponentially with the number of representation features. Thus, for high-dimensional feature spaces, exhaustive SV calculations become infeasible and require the SHAP approximation.

Shapley value-expressed Tanimoto similarity

The SVETA approach has been devised to enable the calculation of exact SVs for SVM models with the Tanimoto kernel (Ralaivola et al., 2005) and binary feature spaces of any dimensionality. It is based on SVs accounting for Tanimoto similarity (Tanimoto, 1958; Flower, 1998) of a predicted instance and each support vector. SVETA values are calculated efficiently by aggregating coalitions with the same numerical value and omitting features having no impact on the similarity.

Tanimoto similarity Sim_{TN}) of two compounds represented by a set of features is defined as the number of shared features, or intersection (I), divided by the count of features present in at least one of the compared instances (union, U). For the following calculations, the union is expressed as the sum of the intersection and symmetric difference (D), ensuring non-overlapping feature categories:

$$Sim_{TN} = \frac{I}{U} = \frac{I}{I+D} \quad \text{(Equation 2)}$$

Equation 2 shows that only the counts of shared and distinct features of the compared instances determine the similarity value. Features belonging to the same category make identical contributions to Sim_{TN}. As a consequence, the corresponding SVs are equal. Accordingly, it is sufficient to calculate the SV for one representative intersecting feature ($f+$) and the SV for a feature from the symmetric difference ($f-$). Notably, features not present in either instance have no effect on the similarity value, resulting in an SV of 0.

The change in Tanimoto similarity that a feature causes when it is added to a coalition of i intersecting features and d features from the symmetric difference must be calculated separately for both cases according to Equations 3 and 4, respectively:

$$\Delta v_{f+}(i,d) = \frac{i+1}{i+d+1} - \frac{i}{i+d} \quad \text{(Equation 3)}$$

$$\Delta v_{f-}(i,d) = \frac{i}{i+d+1} - \frac{i}{i+d} \quad \text{(Equation 4)}$$

Both equations are invalid for empty coalitions ($i + d = 0$) because Tanimoto similarity is undefined for instances containing no feature. Therefore, the numerical value for an empty coalition (represented by the subtracted term) is set to 0, conforming with the SV formalism. According to Equations 3 and 4, coalitions are fully represented by the number of intersecting and distinct features. Hence, SVs can be efficiently calculated only based upon unique combinations of i and d that are then multiplied by the count of their occurrences. The number of possible combinations (C) of i and d also depends on the feature category:

$$C_{f+}(i,d) = \binom{I-1}{i}\binom{D}{d} \quad \text{(Equation 5)}$$

$$C_{f-}(i,d) = \binom{I}{i}\binom{D-1}{d} \quad \text{(Equation 6)}$$

The values of I and D are reduced by one in Equations 5 and 6, respectively, as an assessed feature ($f+$ or $f-$) is not a part of the coalitions.

Finally, the SVs for $f+$ and $f-$are calculated as the sum of the products of Δv, C, and the multinomial coefficient over all unique combinations of i and d:

$$SV_{f+} = \sum_{i=0}^{I-1}\sum_{d=0}^{D}\left(\frac{i+1}{i+d+1} - \frac{i}{i+d}\right)\binom{I-1}{i}\binom{D}{d}\left(\frac{(i+d)!(I+D-i-d-1)!}{(I+D)!}\right) \quad \text{(Equation 7)}$$

$$SV_{f-} = \sum_{i=0}^{I}\sum_{d=0}^{D-1}\left(\frac{i}{i+d+1} - \frac{i}{i+d}\right)\binom{I}{i}\binom{D-1}{d}\left(\frac{(i+d)!(I+D-i-d-1)!}{(I+D)!}\right) \quad \text{(Equation 8)}$$

Exemplary calculation

We consider two binary vectors (comprising five pre-defined features) representing two instances that share two features (set to 1), contain a unique feature each (set to 1 and 0, respectively), and lack a feature (set to 0):

$$x = (1\,0\,0\,1\,0)$$
$$y = (1\,0\,1\,1\,1) \quad \text{(Equation 9)}$$

Table 1. Unique combinations of *i* and *d* for the calculation of Shapley values for features from the intersection

i	d	*v(S)*	*v(SU f+)*	Δ*v*	Number of Coalitions	Inverse multinomial factor
0	0	0	1 / 1 = 1.00	1	1 · 1 = 1	0! · 3! /4! = 0.25
0	1	0 / 1 = 0.00	1 / 2 = 0.50	0.5	1 · 2 = 2	1! · 2! / 4! = 0.08
0	2	0 / 2 = 0.00	1 / 3 = 0.33	0.33	1 · 1 = 1	2! · 1! / 4! = 0.08
1	0	1 / 1 = 1.00	2 / 2 = 1.00	0	1 · 1 = 1	1! · 2! / 4! = 0.08
1	1	1 / 2 = 0.50	2 / 3 = 0.66	0.16	1 · 2 = 2	2! · 1! / 4! = 0.08
1	2	1 / 3 = 0.33	2 / 4 = 0.50	0.16	1*1 = 1	3! · 0! / 4! = 0.25

Also reported are the output values of the coalition with and without the feature, the output difference, the number of relevant coalitions, and the inverse multinomial factor.

These feature settings (with two common features and one unique feature per vector) result in a Tanimoto similarity of 0.5. Table 1 lists all unique combinations of *i* and *d* for coalitions required for the calculation of the SV of a feature that is present in both instances. In addition, the output values of the coalition with and without the feature, the output difference, the number of represented coalitions, and the inverse multinomial factor are reported.

Multiplying the output change with the number of represented coalitions and the inverse multinomial factor for each row and calculating the sum of all values produces an SV of 0.4305 for features present in both instances. Table 2 lists corresponding values for features from the symmetric difference.

The corresponding calculations yield and SV of −0.1805 for the features from the symmetric difference. Given two intersecting, two unique, and one absent feature (SV = 0), the sum of SVs for all features is equal to the similarity value of 0.5.

Feature contributions for support vector machines

The equations above define the calculation of SVs accounting for the Tanimoto similarity of two instances. In the case of the SVM algorithm (Cortes and Vapnik, 1995), SVs for the distance to the separating hyperplane can be obtained by modifying the underlying equation:

$$dist(x) = b + \sum_{n=0}^{N_V} y_n w_n K(x, V_n) \qquad \text{(Equation 10)}$$

In Equation 10, the class label (y, −1 or 1) of a support vector (V) is scaled by its weight (w) and multiplied with the value of the kernel function (K) compared to the predicted instance. The sum of all support vector evaluations is modified by a bias value (b) and gives the distance from the hyperplane for the predicted instance.

SVs can be obtained by replacing the kernel value with the sum of SVETA values over each feature f (Equation 11). After rearranging the order of summation, the SV of a feature is calculated as the sum of SVETA values for the instance and all support vectors, scaled by the support vector label and weight (Equation 12). The bias is considered an additional feature and its numerical value represents the SV.

$$dist(x) = b + \sum_{n=0}^{N_V} y_n c_n \sum_{f}^{F} SV_{f,n} = b + \sum_{f}^{F} \sum_{n=0}^{N_V} y_n c_n SV_{f,n} = SV_b + \sum_{f}^{F} SV_f \qquad \text{(Equation 11)}$$

$$SV_f = \sum_{n=0}^{N_V} y_n c_n SV_{f,n} \qquad \text{(Equation 12)}$$

Calculating contributions to log odds

Because the distance of an instance to the hyperplane is difficult to rationalize without a point of reference, it is often transformed into a probability of class membership via Platt scaling (Platt, 1999):

$$p(x) = \frac{1}{1 + e^{A \cdot dist(x) + B}} \qquad \text{(Equation 13)}$$

Table 2. Unique combinations of *i* and *d* for the calculation of Shapley values for features from the symmetric difference

i	D	*v(S)*	*v(S∪ f+)*	Δ*v*	Number of coalitions	Inverse multinomial factor
0	0	0	0 / 1 = 0.00	0	1 · 1 = 1	0! · 3! / 4! = 0.25
0	1	0 / 1 = 0.00	0 / 2 = 0.00	0	1 · 1 = 1	1! · 2! / 4! = 0.08
1	0	1 / 1 = 1.00	1 / 2 = 0.50	−0.5	2 · 1 = 2	1! · 2! / 4! = 0.08
1	1	1 / 2 = 0.50	1 /3 = 0.33	−0.16	2 · 1 = 2	2! · 1! / 4! = 0.08
2	0	2 / 2 = 1.00	2 / 3 = 0.66	−0.33	1 · 1 = 1	2! · 1! / 4! = 0.08
2	1	2 / 3 = 0.66	2 / 4 = 0.50	−0.16	1 · 1 = 1	3! · 0! / 4! = 0.25

Additional values are reported as in Table 1.

Platt scaling derives the probability as a sigmoidal function of the distance, weighted by a factor *A*, and shifted by the offset *B* (Equation 13). These data-dependent parameters are obtained from a maximum likelihood estimation. SVs for probability estimations cannot be directly calculated from SVs for the distance from the hyperplane, as introduced above. However, instead of estimating the probability, it is possible to calculate log odds (or logit) of class label membership.

$$logit(p(x)) = \log\left(\frac{p(x)}{1 - p(x)}\right) = -A \cdot dist(x) - B \quad \text{(Equation 14)}$$

$$logit(p(x)) = -(A \cdot SV_b + B) - \sum_f^F A \cdot SV_f \quad \text{(Equation 15)}$$

Accordingly, log odds are obtained as a linear transformation of the distances (Equation 14). Therefore, SVs are also linearly transformed (Equation 15). With the offset, a new feature is introduced to conform with the SV formalism, which can be combined with the bias of the SVM, as discussed above, to obtain a new single feature, the contribution of which is determined as $-(A \cdot SV_b + B)$. This feature contribution does not depend on others in analogy to the "expected value" of the SHAP formalism and is thus here also referred to as an expected value.

Proof-of-concept

To further validate SVETA calculations, we generated a model system comprising vectors of low dimensionality (15 features), for which SVs can be calculated directly (Shapley, 2016). A random number generator was used to obtain 20 such vectors, in which each feature had an independent probability of 50% to be present (set to 1) or absent (set to 0). The vectors were systematically compared in a pairwise manner and SVs were calculated directly to yield Tanimoto similarity values. Then, for each pair, SVs were re-calculated using SVETA. In all cases, the SVs were reproduced (differences were consistently smaller than 10^{-10}).

Tanimoto similarity via support vector-expressed Tanimoto similarity vs. Shapley additive explanations

For the low-dimensional model system, SHAP values were also calculated to explain Tanimoto similarities, in each case using the 18 remaining vectors as a background sample. Here, larger deviations were observed. Based on the Fisher transformation (Fisher, 1915), a mean Pearson's correlation coefficient (Baldi et al., 2000) of 0.82 ± 0.25 was obtained for SVs of similarity values calculated using SHAP and SVs, respectively.

Furthermore, pairwise similarity calculations were also carried out for 50 randomly selected dataset compounds represented by Morgan fingerprint features. For this higher-dimensional feature space, SVs can no longer be calculated directly. Compound similarity explanations using SVETA SV and SHAP values (with 48 remaining vectors as a background sample), respectively, yielded a mean correlation coefficient of 0.65 ± 0.29, indicating that SHAP was not well-suited to express Tanimoto similarities. This observation could be rationalized by considering details of the SHAP approach. Following SHAP, features not selected for a coalition are masked by replacing their value with a value from the random background sample. As the frequency of occurrence of features differs, different masking values are selected for a feature that is present

Table 3. Median Person's correlation coefficient values of feature contributions from different models and explanations

	SVs SVM	SHAP values SVM	SVs RF	SHAP values RF
SVs SVM	1	0.682	0.324	0.316
SHAP values SVM	0.682	1	0.634	0.630
SVs RF	0.324	0.634	1	0.996
SHAP values RF	0.316	0.630	0.996	1

or absent. These masking values are expected to influence final values for coalitions in different ways than omitting a non-contributing feature entirely from the evaluation, as in SVETA.

For the output of an SVM model, the similarity between a test instance and learned support vectors plays an essential role. Hence, for an SVM model with Tanimoto kernel, SVETA should be a preferred approach.

Feature contributions

For an adenosine receptor ligand dataset, an SVM classifier was built. In addition, an RF model was derived using the same data. The classification performance of both models was very high, reaching a balanced accuracy of 93 and 92% for SVM and RF, respectively. For the SVM classifier, SHAP values and exact SVETA SVs were calculated. Correspondingly, SHAP values and exact SVs from the TreeExplainer algorithm were computed for the RF model. On the basis of these values, feature contributions to correct compound activity predictions were determined and compared. Median Person's correlation coefficient values of feature contributions of the different models are reported in Table 3 and the value distributions are displayed in Figure 1. For SVM, feature contributions quantified on the basis of SVETA SVs and SHAP values yielded a median correlation coefficient of 0.682, indicating significant differences between prioritized features. Hence, SVM model interpretation on the basis of approximate SHAP values would be limited. By contrast, for the RF model, SHAP and TreeExplainer SVs, a median Person's correlation coefficient close to 1 was obtained. Hence, for RF models, features prioritized on the basis of exact and approximate values were very closely corresponding and model interpretation on the basis of SHAP values would be reliable.

Correlations of feature importance values from different explanatory approaches and models (for example, SHAP values from SVM and SVs from RF) were lower, ranging from 0.316 to 0.634. Limited correlation was at least partly owing to the calculation of different scores for SVM (log odds) and RF (probability). However, the median correlation between SHAP values for SVM and RF reached 0.630, whereas the correlation between exact SVs calculated with SVETA and TreeExplainer, respectively, was much lower (0.324). Hence, SVM and RF prioritized different features for predictions. In the case of RF, the SHAP value provided a highly accurate approximation of exact SVs. By contrast, for SVM, SHAP values had limited accuracy. Taken together, these findings clearly indicated the need for SVM model interpretation on the basis of SVETA values.

Model explanations

For SVM and RF, model-specific explanations of predictions were further analyzed and complemented by feature mapping on test compounds. Therefore, SVETA SVs and TreeExplainer SVs were considered. First, cumulative SVs for present and absent features were calculated over all correctly predicted test compounds. Figure 2 shows the value distributions for active and random test instances. For both SVM and RF models, features present in active compounds determined their correct prediction, whereas the absence of these features was largely responsible for the correct prediction of random instances by RF. These results paralleled prior findings in multi-target activity predictions using RF and TreeExplainer (Feldmann et al., 2021; Feldmann and Bajorath, 2022). However, the comparison also revealed model-specific differences between the highly accurate SVM and RF classifiers. Although features absent in active compounds made only small contributions to RF predictions, they opposed correct SVM predictions. Furthermore, for the SVM model, SVs for random compounds were comparably small for present and absent features. Present features made small negative contributions (supporting correct predictions) while absent features made small positive contributions (opposing the prediction). By contrast, correct predictions of random compounds by the RF model were clearly driven by feature absence, while present features

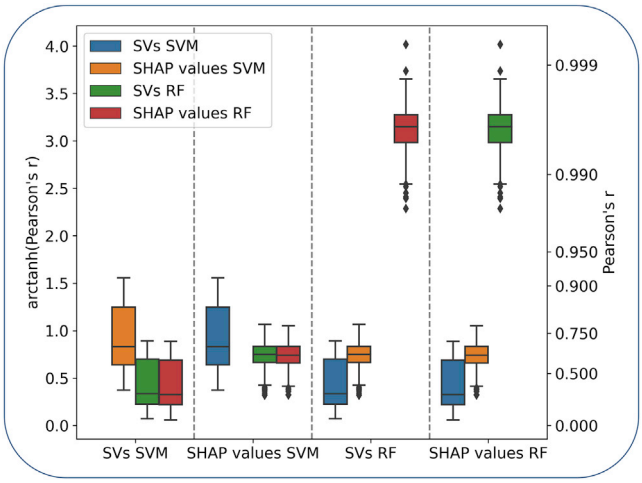


Figure 1. Feature contributions to explanations of predictions with different models

Boxplots report distributions of Pearson's correlation coefficient for feature contributions to correct compound predictions using the SVM and RF model. Depending on the approach, feature contributions were quantified using SHAP values, SVETA SVs, and/or TreeExplainer SVs. For the SVM model, log-odds scores of predictions are explained and for the RF model, class label probabilities.

made essentially no contributions. Hence, while correct predictions of active compounds were similarly determined by SVM and RF, prediction characteristics of random instances differed, despite comparably high prediction accuracy. With respect to random test instances, RF was the more robust classifier, as revealed by SV analysis. For the SVM model, the average sum of all SVs for random compounds was calculated as -0.36 ± 0.90, indicating that correct predictions of these compounds were largely determined by the feature-independent contribution of the expected value (-2.4). On the other hand, the expected value of 0.56 for the class probability of the RF model revealed that this model, without any additional feature information, would generally slightly favor the prediction of active compounds.

SVs of present features were then mapped to exemplary active test compounds (Figure 3). These test compounds represented different chemical series, shown at the top and bottom, respectively. Features making strong positive contributions to the correct prediction of activity prioritized by the SVM and RF model delineated coherent and similar substructures in compounds from both series. Substructures responsible for positive contributions to predicted activity included the core ring system in both compounds. These substructures were reminiscent of adenosine, which represents the endogenous ligand of receptor. By contrast, small negative contributions opposing the correct predictions were mostly centered on individual

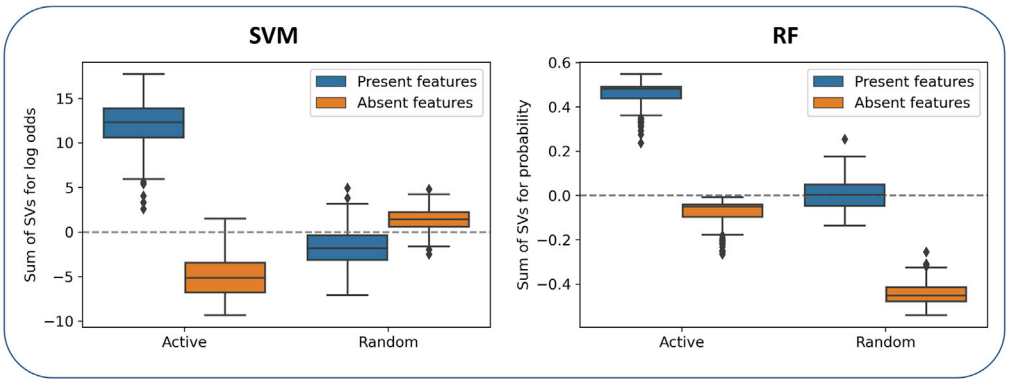


Figure 2. Cumulative contributions of features present or absent in test compounds

Boxplots report SVs for feature contributions to log-odds scores of the SVM model and predicted class probabilities of the RF model.

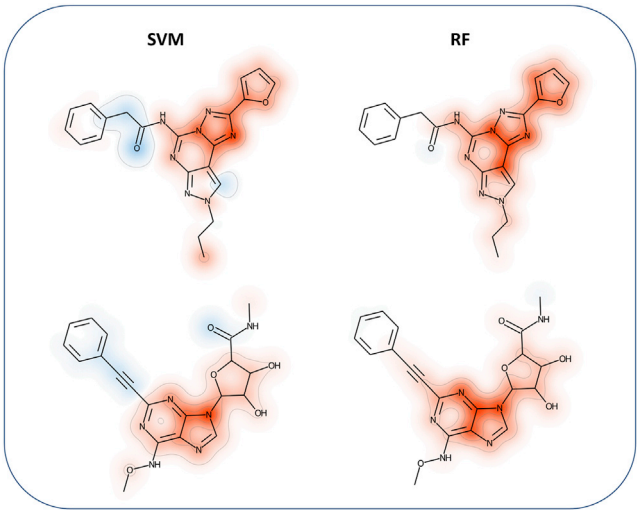


Figure 3. Mapping of feature contributions

SVs of features present in active compounds correctly predicted by the SVM (left) and RF (right) models are mapped on atoms. The atom-based sum of feature contributions is color-coded, with red and blue indicating positive and negative contributions to the prediction of activity, respectively.

atoms distributed over the structures. Hence, the explanations of these exemplary test compounds on the basis of exact SVs derived for the SVM model (SVETA) and RF model (TreeExplainer) were consistent and chemically meaningful.

DISCUSSION

Given the extensive use of ML in many scientific fields and the black box nature of most algorithms, approaches for the explanation of models and their predictions experience increasing attention, especially at interfaces between computation and experiment. A variety of approaches are being considered for XML. Among these is the SV concept originating from game theory. Pioneering work by the Lundberg et al has adapted and further extended this concept for XML. Key aspects of SV analysis include its model-agnostic nature and the ability to quantify contributions of features that are present or absent in test instances to predictions. This sets the approach apart from feature weighting techniques developed earlier and renders it more informative. The SHAP and TreeExplainer framework developed by Lundberg et al. has recently also been introduced in chemoinformatics and drug discovery, primarily focusing on compound property predictions.

In this work, we have reported the development of a new methodology for calculating exact SVs for expressing Tanimoto similarity, which governs different facets of molecular similarity analysis and its applications in chemoinformatics. So far, the calculation of exact SVs has only been feasible for decision tree methods via TreeExplainer. Given its characteristics, Tanimoto similarity is difficult to explain using the local SHAP approximation, as shown herein. By contrast, the newly developed SVETA approach is capable of exactly expressing Tanimoto similarity values. SVETA has been designed to explain predictions of SVMs with Tanimoto kernel. This SVM architecture using binary structural fingerprints as molecular descriptors is a mainstay for compound classification in computer-aided drug discovery, together with decision tree methods. Hence, for molecular ML, SVETA fills a void as it enables an accurate assessment of feature contributions to SVM predictions and the explanation of Tanimoto similarity relationships in predictive modeling. In our proof-of-concept investigation, SVETA produced promising results. The analysis also uncovered similarities and differences in the correct classification of active and random compounds by SVM and RF classifiers, and mapping of contributing features identified based upon SVs produced chemically intuitive results. As a part of our study, the SVETA approach is made freely available. It is hoped that the methodology will spark the interest of many investigators in molecular informatics and drug discovery.

Limitations of the study

The SVETA methodology is currently limited to the use of binary molecular fingerprint descriptors for SVM and the Tanimoto kernel.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Compounds and activity data
  - Molecular representation and activity predictions
  - Model explanations

ACKNOWLEDGMENTS

The authors thank members of the Life Science Informatics groups for scientific discussions.

AUTHOR CONTRIBUTIONS

Conceptualization, C.F. and J.B.; Methodology, C.F. and J.B.; Software, C.F.; Formal Analysis, C.F. and J.B.; Investigation, C.F.; Data Curation, C.F.; Writing – Original Draft, C.F. and J.B.; Writing – Review & Editing, C.F. and J.B.; Visualization, C.F.; Supervision, J.B.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: July 3, 2022
Revised: August 9, 2022
Accepted: August 20, 2022
Published: September 16, 2022

REFERENCES

Baell, J.B., and Holloway, G.A. (2010). New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. J. Med. Chem. *53*, 2719–2740. https://doi.org/10.1021/jm901137j.

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C.A.F., and Nielsen, H. (2000). Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics *16*, 412–424. https://doi.org/10.1093/bioinformatics/16.5.412.

Balfer, J., and Bajorath, J. (2015). Visualization and interpretation of support vector machine activity predictions. J. Chem. Inf. Model. *55*, 1136–1147. https://doi.org/10.1021/acs.jcim.5b00175.

Belle, V., and Papantonis, I. (2021). Principles and practice of explainable machine learning. Front. Big Data *4*, e688969. https://doi.org/10.3389/fdata.2021.688969.

Bento, A.P., Gaulton, A., Hersey, A., Bellis, L.J., Chambers, J., Davies, M., Krüger, F.A., Light, Y., Mak, L., McGlinchey, S., et al. (2014). The ChEMBL bioactivity database: an update. Nucleic Acids Res. *42*, D1083–D1090. https://doi.org/10.1093/nar/gkt1031.

Boser, B.E., Guyon, I.M., and Vapnik, V.N. (1992). A training algorithm for optimal margin classifiers. In Proceedings of the Fifth Annual Workshop on Computational Learning Theory (Association for Computing Machinery), pp. 144–152. https://doi.org/10.1145/130385.130401.

Breiman, L. (2001). Random forests. Mach. Learn. *45*, 5–32. https://doi.org/10.1023/A:1010933404324.

Bruns, R.F., and Watson, I.A. (2012). Rules for identifying potentially reactive or promiscuous compounds. J. Med. Chem. *55*, 9763–9772. https://doi.org/10.1021/jm301008n.

Castelvecchi, D. (2016). Can we open the black box of AI? 2016. Nature *538*, 20–23. https://doi.org/10.1038/538020a.

Chen, H., Engkvist, O., Wang, Y., Olivecrona, M., and Blaschke, T. (2018). The rise of deep learning in drug discovery. Drug Discov. Today *23*, 1241–1250. https://doi.org/10.1016/j.drudis.2018.01.039.

Cortes, C., and Vapnik, V. (1995). Support-vector networks. Mach. Learn. *20*, 273–297. https://doi.org/10.1007/BF00994018.

Feldmann, C., and Bajorath, J. (2022). Differentiating inhibitors of closely related protein kinases with single- or multi-target activity via explainable machine learning and feature analysis. Biomolecules *12*, 557. https://doi.org/10.3390/biom12040557.

Feldmann, C., Philipps, M., and Bajorath, J. (2021). Explainable machine learning predictions of dual-target compounds reveal characteristic structural features. Sci. Rep. *11*, 21594. https://doi.org/10.1038/s41598-021-01099-4.

Fisher, R.A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. Biometrika *10*, 507–521. https://doi.org/10.2307/2331838.

Flower, D.R. (1998). On the properties of bit string-based measures of chemical similarity. J. Chem. Inf. Comput. Sci. *38*, 379–386. https://doi.org/10.1021/ci970437z.

Heikamp, K., and Bajorath, J. (2014). Support vector machines for drug discovery. Expert Opin.

Drug Discov. *9*, 93–104. https://doi.org/10.1517/17460441.2014.866943.

Irwin, J.J., Duan, D., Torosyan, H., Doak, A.K., Ziebart, K.T., Sterling, T., Tumanian, G., and Shoichet, B.K. (2015). An aggregation advisor for ligand discovery. J. Med. Chem. *58*, 7076–7087. https://doi.org/10.1021/acs.jmedchem.5b01105.

Landrum, G., Tosco, P., Kelly, B., Rodríguez, R., Riniker, S., Gedeck, P., Vianello, R., Schneider, N., Kawashima, E., Dalke, A., et al. (2022). RDKit: Open-Source Cheminformatics. http://www.rdkit.org. https://doi.org/10.5281/zenodo.6605135.

Lavecchia, A. (2015). Machine-learning approaches in drug discovery: methods and applications. Drug Discov. Today *20*, 318–331. https://doi.org/10.1016/j.drudis.2014.10.012.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. Nature *521*, 436–444. https://doi.org/10.1038/nature14539.

Lo, Y.-C., Rensi, S.E., Torng, W., and Altman, R.B. (2018). Machine learning in chemoinformatics and drug discovery. Drug Discov. Today *23*, 1538–1546. https://doi.org/10.1016/j.drudis.2018.05.010.

Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. Nat. Mach. Intell. *2*, 56–67. https://doi.org/10.1038/s42256-019-0138-9.

Lundberg, S.M., and Lee, S.-I. (2017). A unified approach to interpreting model predictions. Adv. Neural Inf. Process. Syst. *30*.

Platt, J.C. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In Advances in Large Margin Classifiers (MIT Press), pp. 61–74.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: machine learning in python. J. Mach. Learn. Res. *12*, 2825–2830.

Ralaivola, L., Swamidass, S.J., Saigo, H., and Baldi, P. (2005). Graph kernels for chemical informatics. Neural Netw. *18*, 1093–1110. https://doi.org/10.1016/j.neunet.2005.07.009.

Ribeiro, M.T., Singh, S., and Guestrin, C. (2016). Why should I trust you?": explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (Association for Computing Machinery), pp. 1135–1144.

Rogers, D., and Hahn, M. (2010). Extended-connectivity fingerprints. J. Chem. Inf. Model. *50*, 742–754. https://doi.org/10.1021/ci100050t.

Rodríguez-Pérez, R., and Bajorath, J. (2021). Explainable machine learning for property predictions in compound optimization. J. Med. Chem. *64*, 17744–17752. https://doi.org/10.1021/acs.jmedchem.1c01789.

Sellwood, M.A., Ahmed, M., Segler, M.H., and Brown, N. (2018). Artificial intelligence in drug discovery. Future Med. Chem. *10*, 2025–2028. https://doi.org/10.4155/fmc-2018-0212.

Shapley, L.S. (2016). 17. A value for n-person games. In Contributions to the Theory of Games, H.W. Kuhn and A.W. Tucker, eds. (Princeton University Press), pp. 307–318.

Štrumbelj, E., and Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. Knowl. Inf. Syst. *41*, 647–665. https://doi.org/10.1007/s10115-013-0679-x.

Sun, H., Nguyen, K., Kerns, E., Yan, Z., Yu, K.R., Shah, P., Jadhav, A., and Xu, X. (2017). Highly predictive and interpretable models for PAMPA permeability. Bioorg. Med. Chem. *25*, 1266–1276. https://doi.org/10.1016/j.bmc.2016.12.049.

Tanimoto, T.T. (1958). An Elementary Mathematical Theory of Classification and Prediction (IBM Report).

Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., Li, B., Madabhushi, A., Shah, P., Spitzer, M., and Zhao, S. (2019). Applications of machine learning in drug discovery and development. Nat. Rev. Drug Discov. *18*, 463–477. https://doi.org/10.1038/s41573-019-0024-5.

Ustün, B., Melssen, W.J., and Buydens, L.M.C. (2007). Visualisation and interpretation of support vector regression models. Anal. Chim. Acta *595*, 299–309. https://doi.org/10.1016/j.aca.2007.03.023.

Varnek, A., and Baskin, I. (2012). Machine learning methods for property prediction in chemoinformatics: quo Vadis? J. Chem. Inf. Model. *52*, 1413–1437. https://doi.org/10.1021/ci200409x.

Willett, P. (2014). The calculation of molecular structural similarity: principles and practice. Mol. Inform. *33*, 403–413. https://doi.org/10.1002/minf.201400024.

Yamashita, F., and Hashida, M. (2004). In silico approaches for predicting ADME properties of drugs. Drug Metab. Pharmacokinet. *19*, 327–338. https://doi.org/10.2133/dmpk.19.327.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Compound activity data	ChEMBL 30	https://doi.org/10.6019/CHEMBL.database.30
Confirmed aggregators	Aggregator advisor	http://advisor.docking.org/faq/#Data
Data sets	This paper	https://doi.org/10.17632/hz3pjthz2t.1
Software and algorithms		
RDKit	Zenodo	https://doi.org/10.5281/zenodo.6605135
Lilly Medchem rules	GitHub	https://github.com/IanAWatson/Lilly-Medchem-Rules
Scikit-learn	GitHub	https://github.com/scikit-learn/scikit-learn
Python code for SVETA calculations and analysis workflows	This paper	https://doi.org/10.5281/zenodo.6792073

RESOURCE AVAILABILITY

Lead contact

Further information and requests for code and resources should be directed to and will be fulfilled by the lead contact, Jürgen Bajorath (bajorath@bit.uni-bonn.de).

Materials availability

This study did not generate new unique reagents.

Data and code availability

Compound data have been deposited at Mendeley Data and are publicly available as of the date of publication. Accession numbers are listed in the key resources table.

All original code has been deposited at Zenodo and is publicly available as of the date of publication. DOIs are listed in the key resources table.

Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

Compounds and activity data

Compounds from ChEMBL release 30 (Bento et al., 2014) with standard potency measurements (K_i, IC_{50}, or K_d) and numerically specified potency values (" = ") of at least 10 μM were extracted (and recorded as negative decadic logarithmic values). Compounds with measurements flagged as "potential author error'" or "potential transcription error" were omitted. Furthermore, only activity annotations corresponding to direct interactions (target relationship type: "D) with human wild-type proteins at the highest confidence level (target confidence score: 9) were considered. Molecular mass was calculated using RDKit (Landrum et al., 2022), discarding compounds with a mass of 1000 Da or more. The remaining compounds were computationally screened for colloidal aggregators as identified by the aggregation advisor (Irwin et at., 2015). Such aggregators non-specifically precipitate proteins in assays. In addition, compounds were examined using a substructure-based filter for pan-assay interference compounds (PAINS) (Baell and Holloway, 2010). PAINS are assay interference compounds potentially causing false-positive assay signals by a variety of mechanisms such as absorption of light at probed wavelengths, reactivity, or covalent and unspecific protein binding. Furthermore, the Lilly Medicinal Chemistry filter for potentially reactive and non-specific assay interference compounds was applied, which consists of 275 empirical detection rules derived from screening campaigns and medicinal chemistry knowledge (Bruns and Watson, 2012). For our analysis, a dataset containing 287 compounds with activity against the adenosine receptor A3 (UniProt ID: P0DMS8),

representing active/positive instances, and an equal number of randomly selected ChEMBL compounds (assumed inactive/negative) was generated. From this dataset, a random subset of 50 compounds was taken for similarity control calculations.

Molecular representation and activity predictions

Compounds were represented as binary bit vectors, in which each element corresponded to a unique structural feature generated with the Morgan algorithm using a bond radius of 2 (Rogers and Hahn, 2010), as implemented in RDKit.

An RF classifier comprising an ensemble of decision trees (Breiman, 2001; Pedregosa et al., 2011) and an SVM model were generated based on a training set comprising 50% of the (positive/negative) compounds. The remaining 50% were used as the test set. Optimal hyperparameters (SVM: C∈{0.1, 1, 10, 50, 100, 200, 400, 500, 750, 1000}; RF: n_estimators∈{10, 100, 250, 500}, min_samples_split∈{2, 3, 5, 7, 10}, min_samples_leaf∈{1, 2, 5, 10}) were determined over 10 internal random training-validation splits (50/50) and applied to derive final classifiers for the complete training set. Model performance on the test set was quantified as balanced accuracy (BA) defined as:

$$BA = \frac{1}{2} \cdot \left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP} \right) \quad \text{(Equation 16)}$$

In Equation 16, *TP*, *FP*, *TN*, and *FN* represent the number of true positives, false positives (*FP*), true negatives, and false negatives, respectively.

Model explanations

RF and SVM predictions were analyzed using the SHAP kernel explainer producing locally approximated feature importance values. As a background sample for SHAP, 100 randomly selected training set compounds were used. In addition, for RF, the TreeExplainer algorithm (Lundberg et al., 2020) with interventional feature perturbation using the entire training set as a background sample was applied to calculate exact importance values. Furthermore, SVM predictions were explained with the newly introduced SVETA approach.