JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

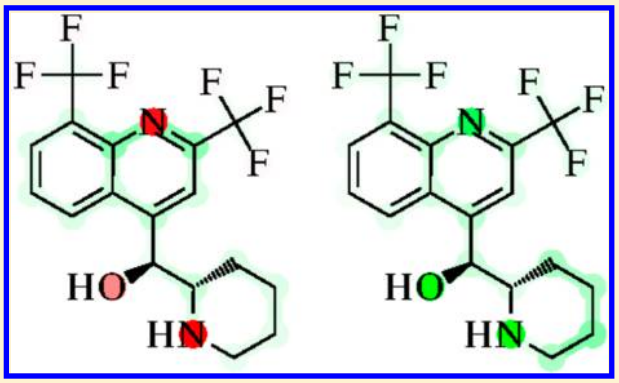
pubs.acs.org/jcim

Visualization and Interpretation of Support Vector Machine Activity Predictions

Jenny Balfer and Jürgen Bajorath*

Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Dahlmannstr. 2, D-53113 Bonn, Germany

ABSTRACT: Support vector machines (SVMs) are among the preferred machine learning algorithms for virtual compound screening and activity prediction because of their frequently observed high performance levels. However, a well-known conundrum of SVMs (and other supervised learning methods) is the black box character of their predictions, which makes it difficult to understand why models succeed or fail. Herein we introduce an approach to rationalize the performance of SVM models based upon the Tanimoto kernel compared with the linear kernel. Model comparison and interpretation are facilitated by a visualization technique, making it possible to identify descriptor features that determine compound activity predictions. An implementation of the methodology has been made freely available.



INTRODUCTION

Support vector machines (SVMs) are among the most widely used machine learning algorithms in chemoinformatics,[1] especially for compound activity prediction. SVMs were originally used for binary object classification (e.g., active vs inactive compounds),[2] but have also been adapted for multitarget predictions[3−6] and compound ranking.[7,8] The popularity of SVMs is due to their ability to reach higher performance levels than other prediction methods in many applications.[1] A foundation of the SVM approach is the use of kernel functions to project data sets into higher-dimensional space representations in which a linear separation of positive and negative training instances is feasible. For ligand-based virtual screening, the Tanimoto kernel[9] is often used in combination with binary fingerprints as molecular representations. The Tanimoto kernel utilizes the well-known Tanimoto similarity formalism[10] and is parameter-free, which renders it attractive for chemoinformatics applications.

A conundrum of the SVM approach (and also other machine learning methods, such as neural networks) is its black box character, which refers to the inability to rationalize why prediction models succeed or fail and interpret them in chemical terms. This also means that it is generally difficult to modify models or molecular representations for specific applications. Only a few attempts to rationalize SVM modeling and performance have been made to date. Previous work on SVM model interpretation has usually focused on linearly separable data,[11,12] thereby avoiding analysis in high-dimensional kernel-dependent reference spaces. In the presence of nonlinear data−property relationships, data were partitioned into several local Voronoi regions prior to SVM modeling, and local SVM models were separately built for each of these regions.[11] The weights of the support vectors from which the models were generated were then used to assess the importance of each chosen molecular descriptor.[13,14] Another approach to assess the importance of descriptors in nonlinear SVMs internally stores information during kernel calculation and then readjusts these weights with linear SVM coefficients.[15] This method is only applicable if feature importance information is available for the given kernel. In addition, partial derivatives of a kernel function were used to identify descriptors with the largest gradient components, which were hypothesized to be the most important for prediction.[16] In this case, only the derivatives of the kernel function need to be provided.

Different from model internal analysis, "rule extraction" from SVMs has also been attempted.[17] In this case, one tries to mimic the classification of an SVM model as closely as possible, without interpreting the model itself, to derive a set of rules approximating SVM classification. Hence, these approaches aim at an indirect assessment of SVM predictions. Rule extraction often suffers from the lack of clear rule definitions and is difficult to apply in high-dimensional reference spaces,[17] a hallmark of SVM modeling.

Following a different approach, Hansen et al.[18] have proposed a method for prediction visualization in which the most important support vectors are displayed together with their factors. This method has the principal advantage that it does not require any prior knowledge about the kernel function used. However, in this case it is not possible to explain the influence of single descriptors or features on SVM classification. For this purpose, "explanation vectors" representing local gradients of input descriptors are derived. Therefore, the SVM model must also be mimicked by another classifier such as

Received: March 31, 2015
Published: May 19, 2015

DOI: 10.1021/acs.jcim.5b00175
J. Chem. Inf. Model. 2015, 55, 1136−1147

Parzen windows.[19] The utility of such explanation vectors typically depends on the data sets under study.

Herein we introduce a new methodology for visualization and interpretation of SVM predictions using the Tanimoto kernel in comparison with the linear kernel. It provides intuitive access to descriptor features that are the most important for a given SVM prediction and enables mapping of features onto molecular graphs. An interactive graphical user interface is employed for visualization. The methodology clearly reveals how the Tanimoto kernel facilitates many accurate activity predictions and rationalizes failures of the linear kernel.

CONCEPTS AND METHODS

Support Vector Machine Theory. SVMs aim to solve a classification task by finding a hyperplane in feature space that best separates training examples having different binary class labels.[20] Test instances are then classified on the basis of the side of the separating hyperplane on which they fall, as determined by the following decision function:

$$f(\mathbf{x}) = \mathrm{sign}(\langle \mathbf{w}, \mathbf{x} \rangle - b) \tag{1}$$

where $\mathbf{w}$ is the normal vector of the separating hyperplane, $\mathbf{x}$ is the test instance, and b is the so-called bias of the hyperplane. If $\{(\mathbf{x}^{(i)}, y^{(i)}) \mid i = 1, ..., n\}$ is a set of n training examples $\mathbf{x}^{(i)}$ with known class labels $y^{(i)} \in \{-1, +1\}$, the parameters $\mathbf{w}$ and b of the hyperplane are derived by solving the following optimization problem:[20]

$$\min \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^{n} \xi^{(i)} \tag{2}$$

subject to

$$y^{(i)}(\langle \mathbf{w}, \mathbf{x}^{(i)} \rangle - b) \geq 1 - \xi^{(i)}, \quad i \in \{1, ..., n\}$$
$$\xi^{(i)} \geq 0, \quad i \in \{1, ..., n\} \tag{3}$$

Minimizing $\mathbf{w}$ yields the hyperplane with the maximum distance to training examples on either side, the so-called margin. The parameter C needs to be adjusted to control the balance between correct classification of training examples and permitted prediction errors, which is of critical importance for model generalization. Misclassifications are represented by the slack variables $\xi^{(i)}$ introduced to allow a certain number of training errors.[21]

Instead of directly solving the primal optimization problem, it is also possible to formulate an equivalent dual problem using Lagrangian multipliers.[20] In this formulation, the constraints of the original problem are added to the objective function, and the resulting dual problem must be optimized. By application of the Karush−Kuhn−Tucker conditions,[22] it is possible to formulate the dual problem for the primal optimization problem in eq 2 as follows:

$$\max\left(\sum_{i=1}^{n} \lambda^{(i)} - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda^{(i)} \lambda^{(j)} y^{(i)} y^{(j)} \langle \mathbf{x}^{(i)}, \mathbf{x}^{(j)} \rangle \right) \tag{4}$$

subject to

$$\sum_{i=1}^{n} \lambda^{(i)} y^{(i)} = 0$$
$$0 \leq \lambda^{(i)} \leq C, \quad i \in \{1, ..., n\} \tag{5}$$

where $\lambda^{(i)}$ are the Lagrangian multipliers that are introduced when the constraints of the primal problem are embedded into the objective function of the dual problem. This formulation makes it possible to compute the normal vector of the hyperplane as

$$\mathbf{w} = \sum_{i=1}^{n} \lambda^{(i)} y^{(i)} \mathbf{x}^{(i)} \tag{6}$$

Since the Lagrangian multipliers $\lambda^{(i)}$ can be nonzero only for training examples that lie on or in the margin of the hyperplane or are misclassified, it is also possible to reduce eq 6 to this subset of training examples, the so-called support vectors.[20] Hence, the majority of training examples can be discarded following the training phase, which makes SVM modeling suitable for large data sets.

Another advantage of the dual formulation is that it enables the application of the "kernel trick".[23] The underlying idea is that data that cannot be linearly separated in the original feature space are projected into a higher-dimensional kernel space in which linear separation might become feasible. In this case, the normal vector $\mathbf{w}$ has the higher dimensionality of the kernel space, and training examples are projected into this new space via a mapping function $\phi(\mathbf{x})$. This changes only the constraints in eq 3:

$$y^{(i)}(\langle \mathbf{w}, \phi(\mathbf{x}^{(i)}) \rangle - b) \geq 1 - \xi^{(i)}, \quad i \in \{1, ..., n\} \tag{7}$$

Analogously, the dot product of the examples in eq 4 is replaced by the dot product of their mappings:

$$\max\left(\sum_{i=1}^{n} \lambda^{(i)} - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda^{(i)} \lambda^{(j)} y^{(i)} y^{(j)} \langle \phi(\mathbf{x}^{(i)}), \phi(\mathbf{x}^{(j)}) \rangle \right) \tag{8}$$

Using Mercer's theorem,[24] we can replace the dot product in the dual objective function by a kernel function $K(u, v)$ that implicitly computes $\langle \phi(\mathbf{x}^{(i)}), \phi(\mathbf{x}^{(j)}) \rangle$ without explicitly mapping $\mathbf{x}^{(i)}$ and $\mathbf{x}^{(j)}$ into the high-dimensional kernel space. To compute the normal vector of the hyperplane in kernel space, one would need to apply the explicit mapping $\phi(\mathbf{x})$:

$$\mathbf{w} = \sum_{\text{support vectors}} \lambda^{(i)} y^{(i)} \phi(\mathbf{x}^{(i)}) \tag{9}$$

In practice, this explicit derivation of $\mathbf{w}$ is not required because the decision function can also be expressed using the kernel:

$$f(x) = \mathrm{sign}(\langle \mathbf{w}, \phi(\mathbf{x}) \rangle - b)$$
$$= \mathrm{sign}\left(\sum_{\text{support vectors}} \lambda^{(i)} y^{(i)} K(\mathbf{x}^{(i)}, \mathbf{x}) - b \right) \tag{10}$$

This procedure enables the use of kernel spaces that are theoretically infinite because the mapping functions do not need to be computed explicitly. A variety of kernel functions have been developed, including the Gaussian or radial basis function kernel, the Tanimoto kernel, and more complex graph kernels.[9,25,26] SVMs utilizing kernels usually have much higher prediction capacity than linear models.[1] However, the use of kernel functions comes at the price of black box character and lack of model interpretability. If the explicit mapping $\phi(\mathbf{x})$ is not available, it is impossible to determine contributions of the features to the classification.

DOI: 10.1021/acs.jcim.5b00175
J. Chem. Inf. Model. 2015, 55, 1136−1147

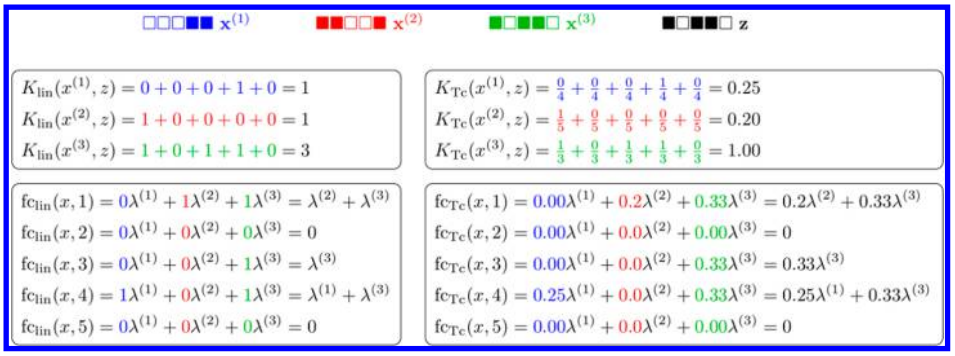


Figure 1. Example calculation. Shown is a minimal example consisting of three support vectors $\mathbf{x}^{(1)}$–$\mathbf{x}^{(3)}$ and one test compound $\mathbf{z}$, represented as fingerprints with five features. Filled and unfilled squares represent features that are set on and off, respectively. The support vectors are colored according to their contributions to the formulas shown below. On the left and right, example kernel and feature contribution calculations are shown for the linear and Tanimoto kernels, respectively.

Feature Weighting. For the linear kernel, feature importance can be easily interpreted because it can be expressed as a sum of individual feature contributions:

$$K_{\text{linear}}(\mathbf{u}, \mathbf{v}) = \langle \mathbf{u}, \mathbf{v} \rangle = \sum_{d=1}^{D} u_d v_d \tag{11}$$

Thus, it is readily possible to weight each feature by $\lambda^{(i)} y^{(i)}$ and calculate the classification function as a sum of weighted feature contributions. However, nonlinear kernels often cannot be expressed as a sum of feature contributions. Nevertheless, they might be modified accordingly. For example, let us consider the Tanimoto kernel, defined as

$$\begin{aligned} K_{\text{Tanimoto}}(\mathbf{u}, \mathbf{v}) &= \frac{\langle \mathbf{u}, \mathbf{v} \rangle}{\langle \mathbf{u}, \mathbf{u} \rangle + \langle \mathbf{v}, \mathbf{v} \rangle - \langle \mathbf{u}, \mathbf{v} \rangle} \\ &= \frac{\sum_{d=1}^{D} u_d v_d}{\langle \mathbf{u}, \mathbf{u} \rangle + \langle \mathbf{v}, \mathbf{v} \rangle - \langle \mathbf{u}, \mathbf{v} \rangle} \\ &= \sum_{d=1}^{D} \frac{u_d v_d}{\langle \mathbf{u}, \mathbf{u} \rangle + \langle \mathbf{v}, \mathbf{v} \rangle - \langle \mathbf{u}, \mathbf{v} \rangle} \end{aligned} \tag{12}$$

Under the condition that the denominator is constant, it is possible to express the Tanimoto kernel as a sum of feature contributions. Since $\mathbf{u}$ and $\mathbf{v}$ are constant for any single kernel calculation, the condition of a constant denominator is applicable in this case. To calculate $\text{fc}(\mathbf{x}, d)$, the contribution of feature d to an individual SVM prediction, the following equations are applied:

$$\text{fc}_{\text{linear}}(\mathbf{x}, d) = \sum_{\text{support vectors}} y^{(i)} \lambda^{(i)} x_d^{(i)} x_d \tag{13}$$

$$\text{fc}_{\text{Tanimoto}}(\mathbf{x}, d) = \sum_{\text{support vectors}} \frac{y^{(i)} \lambda^{(i)} x_d^{(i)} x_d}{\langle \mathbf{x}^{(i)}, \mathbf{x}^{(i)} \rangle + \langle \mathbf{x}, \mathbf{x} \rangle - \langle \mathbf{x}^{(i)}, \mathbf{x} \rangle} \tag{14}$$

The denominator in eq 14 is constant only for each individual support vector. Nonetheless, it is possible to express the feature contribution as a sum. To clarify this point, we consider an exemplary case with three support vectors and five features, as shown in Figure 1. Here the three support vectors are labeled as $\mathbf{x}^{(1)}$, $\mathbf{x}^{(2)}$, and $\mathbf{x}^{(3)}$ and colored blue, red, and green, respectively, while the fingerprint of the compound to be predicted is shown in black and labeled as $\mathbf{z}$. For the linear kernel (left), the derivation is straightforward. The first two support vectors share one bit with the test compound and the third shares three. Accordingly, the feature contributions are derived (lower left). Here only $\lambda^{(i)} x_d^{(i)} x_d$ is shown, and the $y^{(i)}$ have been omitted for clarity. Since the second and fifth bits in the model fingerprint make no contribution to the kernel values, their feature contributions are zero. By contrast, the contributions of the first, third, and fourth bits are derived from the support vectors making a nonzero contribution to the respective sum.

On the right in Figure 1, the same calculations are reported for the Tanimoto kernel. Here the similarities of the support vectors and the test compound are weighted according to eq 12 and are therefore not the same for the first two support vectors. The denominator of each single kernel calculation is constant, but there are different denominators for each support vector. On the lower right, the derivations of the feature contributions are shown. Again, the second and fifth features do not contribute to the final prediction. However, the other three contributions are derived as weighted sums from the support vectors, in which not only $\lambda^{(i)}$ but also the different denominators contribute to the weighting. In the example shown, $\lambda^{(3)}$ has consistently higher weights than $\lambda^{(1)}$ or $\lambda^{(2)}$.

Prediction Visualization. To visualize SVM predictions, we use a graphical method reminiscent of the user interface previously introduced for visualization of naïve Bayesian classification models.[27] Each descriptor feature is visualized as a single point in a polar coordinate system. The more a feature contributes to the prediction, the more remote it is from the pole. Feature points making negative and positive contributions to a prediction are colored red and blue, respectively. Features can be organized into structural subsets displayed in different regions of the polar coordinate system. In the prototypical implementation we provide (see below), feature points can be interactively selected to access associated information. Figure 2 shows an exemplary visualization of a theoretical prediction wherein only three features make nonzero contributions to the prediction. These features include the substructure "NC(O)O" with a contribution of −0.1, the pattern "S=A" (where A refers to any aliphatic atom) with a contribution of 0.3, and the feature "halogen" with a contribution of 0.4. Hence, the final sum of these contributions is 0.6, meaning that this compound would be predicted as active in any model with a bias lower than 0.6. The bias, as defined above, can also be rationalized as a model threshold for the prediction of activity (i.e., if the sum of feature contributions exceeds the bias, a compound is predicted to be active). The relative percentage scale in Figure

DOI: 10.1021/acs.jcim.5b00175
J. Chem. Inf. Model. 2015, 55, 1136−1147

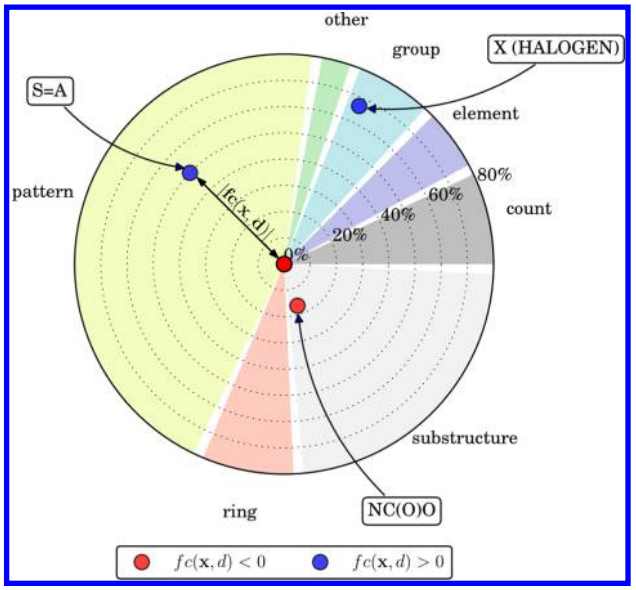


Figure 2. Principles of prediction visualization. Features are shown as points on a polar coordinate system and color-coded by positive (blue) and negative (red) feature contributions. The distance of a feature point from the pole reflects the magnitude of its contribution. Feature points are organized into groups and shown on differently colored backgrounds. A relative scale is provided that gives the percentage contribution of the feature to the overall sum.

2 can be used to easily access the relative importance of each feature. For instance, the halogen feature accounts for almost 67% of the final sum. While the distance of each point from the pole represents the magnitude of a feature contribution, colors refer to positive (blue) or negative (red) absolute contributions.

Feature Mapping. In addition to visualization of SVM predictions, feature contributions are also mapped back onto the molecular graph of the classified compound. For this purpose, we use an approach similar to that of Rosenbaum et al.[12] Each atom and bond in the molecular graph is assigned a weight accounting for its accumulated feature contributions. In the case of fingerprint descriptors, we first determine each feature that is set on and then locate the corresponding substructures in compound **x**. Each participating atom a_x and bond b_x is then assigned a feature contribution, and the contributions of overlapping features are added:

$$w(b_x) = \sum_{\{d|b_x \in d\}} \mathrm{fc}(\mathbf{x}, d) \tag{15}$$

For mapping, feature contributions are normalized with respect to the numbers of atoms and bonds in the corresponding substructures:

$$w(b_x) = \sum_{\{d|b_x \in d\}} \frac{n(d)\mathrm{fc}(\mathbf{x}, d)}{n_{\mathrm{atoms}}(\mathbf{x}, d) + n_{\mathrm{bonds}}(\mathbf{x}, d)} \tag{16}$$

where $n(d)$ is the number of times a substructure must occur in the molecule for feature d to be set. Usually, one occurrence of a substructure is required for a bit to be set on, but there are also features requiring more than one occurrence, such as "more than two nitrogens", for which $n(d) = 3$. This normalization is applied to ensure that atoms and bonds in large or recurrent substructures are assigned only a fraction of the feature weight while single-atom features are fully taken into account in the mapping.

The resulting weights are color-coded such that white corresponds to a weight of zero, red to a negative weight, and green to a positive weight. A weight of zero may occur if a certain atom or bond is not part of any substructure feature encoded by a fingerprint or if the contributions of overlapping features add up to zero. The color shading scales with the magnitude of negative or positive contributions (i.e., the darker the shading, the larger the magnitude).

MATERIALS AND PROTOCOLS

Compound Data Sets. From ChEMBL version 20,[28] three large sets of compounds active against different G-protein-coupled receptors with available high-confidence activity data for individual human targets were extracted, as summarized in Table 1. Compounds were required to be tested in direct binding assays with K_i values of at most 10 000 nM. In-house filters were applied to remove duplicate, highly reactive, and PAINS[29] compounds. Additionally, a set of mitogen-activated protein (MAP) kinase p38 alpha inhibitors was selected using the same criteria as above (Table 1). However, in this case, only IC_{50} measurements were available. Furthermore, 10 000 compounds not contained in these three data sets were randomly extracted from each of ChEMBL version 20 and ZINC version 12[30] and used as negative training examples.

Table 1. Data Sets[a]

TID	target name	no. of compounds
252	adenosine A2a receptor	2646
259	cannabinoid CB2 receptor	2202
72	dopamine D2 receptor	2200
10188	MAP kinase p38 alpha	1485

[a]For each data set, the ChEMBL target ID (TID), the target name, and the number of active compounds is reported.

Molecular Representation. In this study, we used the MACCS substructural fingerprint as a molecular representation.[31] The MACCS fingerprint consists of 166 bits, each of which encodes a predefined substructure or pattern. Fingerprint representations were computed with an in-house implementation using OpenEye's OEChem toolkit[32] and SMARTS patterns adapted from RDKit.[33] For visualization, MACCS features were organized into seven different groups, including "ring", "count", "group", "element", "substructure", "pattern", and "other".[27] This approach is applicable to any fingerprint.

Parameter Selection. First, we divided each data set into a training set containing 80% of its compounds and a test set with the remaining 20%. To select the best regularization term C for SVM models, 10-fold cross-validation was performed on the training set. For this purpose, the training set was randomly divided into 10 equally sized subsets. Each of these subsets was used once as a validation set, and models were built with varying C parameter on the remaining nine subsets. To assess the performance of all intermediate and final models, we used the F_1 score, defined as

$$F_1 = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{17}$$

where TP refers to true positive, FP to false positive, and FN to false negative predictions. Following parameter variation, the value of C yielding the best mean F_1 score across all subsets was selected.

DOI: 10.1021/acs.jcim.5b00175
J. Chem. Inf. Model. 2015, 55, 1136−1147

To account for typically unbalanced SVM classification tasks (i.e., with more inactive than active compounds available), two adjusted versions of C, denoted as C_+ and C_-, can be used to account for the slack variables of positive and negative training examples, respectively:[34]

$$\min \frac{1}{2}\|\mathbf{w}\|^2 + C_+ \sum_{\{i|y^{(i)}=+1\}} \xi^{(i)} + C_- \sum_{\{i|y^{(i)}=-1\}} \xi^{(i)} \quad (18)$$

The terms C_+ and C_- are often derived such that their ratio is equal to the inverse ratio of active and inactive compounds:[34]

$$\frac{C_+}{C_-} = \frac{\text{no. of negative examples}}{\text{no. of positive examples}} \quad (19)$$

During cross-validation, C_- was first varied on a coarse grid of $\{2^x \mid x \in \{-5, -3, ...,15\}\}$ to preselect the best values, which was followed by cross-validation on a finer grid around preselected values of $\{2^x \mid x \in \{C_{best} - 2, C_{best} - 1.75, ..., C_{best} + 2\}\}$, leading to the final selection of C_-.[35] For each value of C_-, two models were trained: one with $C_+ = C_-$ and the other with a C_+ value derived via eq 19. The value combinations giving the best F_1 scores are summarized in Table 2. With one exception, the $C_+ = C_-$ setting was preferred, indicating that a potential influence of data imbalance was mostly negligible here.

Table 2. Model Parameters[a]

TID	kernel	C_-	C_+
252	linear	8.00	8.00
252	Tanimoto	45.25	45.25
259	linear	45.25	45.25
259	Tanimoto	26.91	128.74
72	linear	64.00	64.00
72	Tanimoto	32.00	32.00
10188	linear	0.11	0.11
10188	Tanimoto	64.00	64.00

[a]Given are the best values of C_- and C_+ for each compound set and kernel function as determined via cross-validation (see the text).

Final SVM Models. For each combination of activity class and kernel, the C_+ and C_- values reported in Table 2 were then used to train the final SVM prediction models on 80% of the compounds randomly selected from each activity class. Final model performance was assessed on the basis of F_1 scores derived on the test sets. All of the models were generated using the freely available implementation SVM^{light}.[36]

Software Used and Implementation. For feature mapping and compound display, OpenEye's OEChem and OEDepict toolkits[32,37] were used. Visualizations of predictions were generated using Matplotlib.[38] A prototypical Python implementation of the visualization methodology reported herein was made freely available via the open access platform Zenodo.[39]

RESULTS AND DISCUSSION

Interpretability of machine learning models is of high value for structure−activity relationship analysis.[12,16,18,27,40,41] For this purpose, we introduce an approach for the visualization of individual SVM predictions and identification of key features that determine activity predictions. Initially, as a basis for our investigation, we analyze SVM model performance for exemplary data sets and compare kernels. Then we focus on visualization, feature identification, and mapping.

Model Performance. Figure 3 summarizes the performance of the final SVM models using the ChEMBL subset as

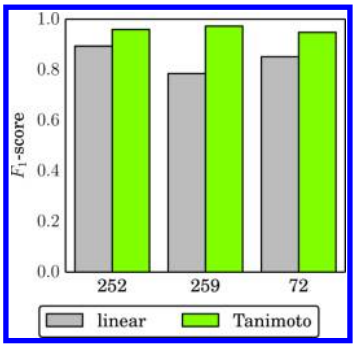


Figure 3. SVM model performance. Reported are the F_1 scores of the final models for the four target sets.

inactive training compounds. Generally high prediction accuracy was observed, with F_1 scores ranging from 78.5% to 97.3%. As anticipated, the use of the Tanimoto kernel led to more accurate prediction than the simple linear kernel, with differences in F_1 scores ranging from 6.6% (adenosine A2a receptor ligands) to 18.8% (cannabinoid CB2 receptor ligands). The models derived using ZINC compounds as inactive training examples yielded overall similar performance, with deviations below 2%. The only exception was the linear model for the cannabinoid CB2 receptor, which performed 7% better using ZINC compounds as inactives than ChEMBL compounds. Hence, unless stated otherwise, we will focus on the models derived using ChEMBL compounds as inactives in the following discussion.

Kernel Comparison. Tables 3−6 report predictions for the four activity classes to compare the two kernels at the level of

Table 3. Predictions for Adenosine A2a Receptor Ligands[a]

		linear			
		TP	FN	TN	FP
Tanimoto	TP	469	38		
	FN	6	8		
	TN			1935	45
	FP			7	22

[a]For the final SVM models, the numbers of true positives (TPs), false positives (FPs), true negatives (TNs), and false negatives (FNs) are reported. The results are presented in a matrix format. For example, there were 469 TPs and 1935 TNs shared by SVM models using the Tanimoto and linear kernels, whereas six FNs using the Tanimoto kernel were TPs using the linear kernel and 45 TNs using the Tanimoto kernel were FPs using the linear kernel.

individual compounds, distinguishing between true positives (TPs), false negatives (FNs), true negatives (TNs), and false positives (FPs). Consistent with the overall high prediction accuracy, most of the compounds were correctly predicted to be active or inactive using both kernels (and are thus reported on the diagonal of the tables). However, for SVM model diagnostics and visualization, compounds yielding different predictions with alternative kernels (represented by off-diagonal numbers in tables) are prime examples.

From the subsets for which the Tanimoto models yielded correct predictions and the linear model incorrect predictions,

DOI: 10.1021/acs.jcim.5b00175
J. Chem. Inf. Model. 2015, 55, 1136−1147

Table 4. Predictions for Cannabinoid CB2 Receptor Ligands[a]

		linear			
		TP	FN	TN	FP
Tanimoto	TP	318	89		
	FN	5	10		
	TN			1937	74
	FP			4	4

[a]For the final SVM models, the numbers of true positives (TPs), false positives (FPs), true negatives (TNs), and false negatives (FNs) are reported. The data representation is according to Table 3.

Table 5. Predictions for Dopamine D2 Receptor Ligands[a]

		linear			
		TP	FN	TN	FP
Tanimoto	TP	384	45		
	FN	3	16		
	TN			1911	53
	FP			7	21

[a]For the final SVM models, the numbers of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) are reported. The data representation is according to Table 3.

Table 6. Predictions for MAP Kinase p38 Alpha Inhibitors[a]

		linear			
		TP	FN	TN	FP
Tanimoto	TP	239	37		
	FN	1	12		
	TN			1962	32
	FP			5	10

[a]For the final SVM models, the numbers of true positives (TPs), false positives (FPs), true negatives (TNs), and false negatives (FNs) are reported. The data representation is according to Table 3.

test compounds were selected having the largest difference of the predictive function:

$$\arg\max_{\mathbf{x}} |f_{\text{linear}}(\mathbf{x}) - f_{\text{Tanimoto}}(\mathbf{x})| \quad (20)$$

These compounds are given in Table 7, and their predictions are analyzed in the following. Table 7 also reports for each compound the sum of feature contributions from each prediction as well the SVM model biases for prediction of activity determined during the training phase. Cumulative feature contributions can be negative or positive. With increasing positive magnitude, the likelihood of positive activity predictions increases. A compound is predicted to be active if the sum of the feature contributions exceeds the model bias.

Table 7. Details of Individual Predictions[a]

		linear		Tanimoto	
TID	CID	$\sum$fc($\mathbf{x}$, d)	b	$\sum$fc($\mathbf{x}$, d)	b
252	CHEMBL411685	5.16	9.15	4.46	3.46
252	CHEMBL11113	12.94	9.15	3.15	3.46
259	CHEMBL1834525	−4.32	2.17	3.45	2.81
259	CHEMBL585041	5.10	2.17	2.15	2.81
72	CHEMBL419792	1.51	5.90	2.89	2.83
72	CHEMBL12028	8.14	5.90	2.12	2.83
10188	CHEMBL320069	2.08	5.74	4.33	3.33
10188	CHEMBL57	7.35	5.74	2.84	3.33

[a]For individual predictions discussed in the text, the ChEMBL target ID (TID), compound ID (CID), sum of feature contributions $\sum$fc($\mathbf{x}$, d), and model bias b for each kernel are reported. If the sum of the feature contributions is larger than the bias, the compound is predicted to be active; otherwise, it is predicted to be inactive.

Visualization of Predictions and Feature Mapping. The graphical analysis of predictions aims to identify descriptor features that make important contributions to correct and incorrect SVM predictions of compound activity using different kernel functions.

Adenosine A2a Receptor Ligands. Figure 4 shows prediction visualizations using the linear and Tanimoto kernels

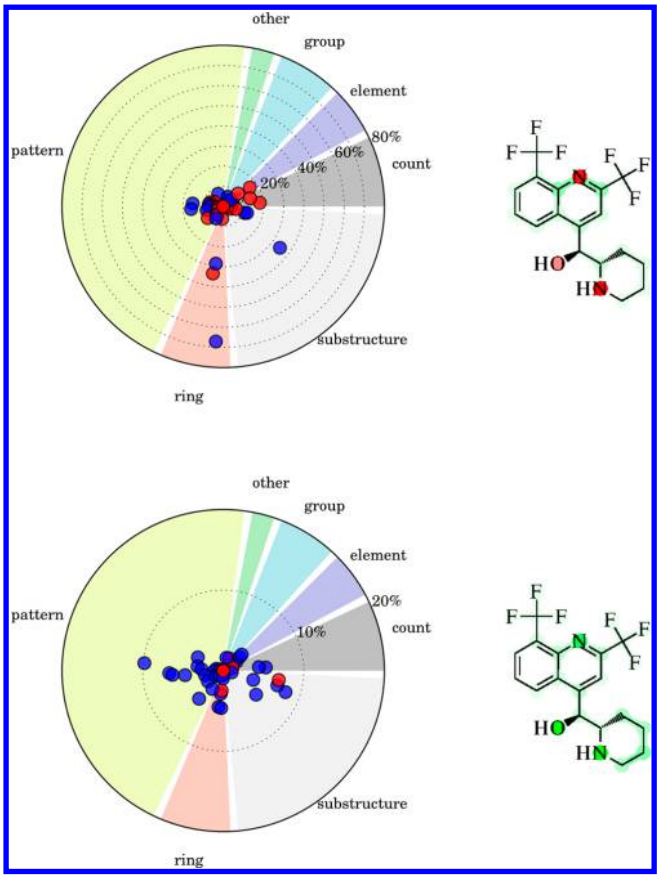


Figure 4. Visualization of predictions and feature mapping for ChEMBL compound 411685. Individual predictions using SVM models with the (top) linear and (bottom) Tanimoto kernels are visualized, and key structural features are mapped. In each panel showing predictions for a given compound, the relative scale of the visualizations has been adjusted such that the magnitudes of feature contributions can be directly compared. This adenosine A2a receptor ligand was predicted to be inactive by the linear and active by the Tanimoto model.

and mappings of MACCS fingerprint features for an active adenosine A2a receptor ligand. The compound was incorrectly predicted to be inactive by the linear kernel but correctly predicted to be active by the Tanimoto kernel. The visualization for the linear model (Figure 4 top) identified one feature with a large positive contribution to the prediction ("six-membered ring"), two features with intermediate positive contributions ("aromatic ring >1" and "C:N", where ":" denotes an aromatic bond), and one feature with an intermediate negative contribution ("N heterocycle"). Despite the preva-

DOI: 10.1021/acs.jcim.5b00175
J. Chem. Inf. Model. 2015, 55, 1136−1147

lence of three features making large or intermediate positive contributions, the linear model yielded a false negative prediction. All of the other MACCS features mapped to the center of the graph, corresponding to small-magnitude contributions. Although the six-membered ring made a feature contribution of almost 70% to the overall sum, the feature mapping shown in Figure 4 revealed that negative contributions of smaller magnitude also occurred at several ring atom positions, which reduced the positive contribution of the six-membered ring. Ultimately, the individual contributions led to a sum of feature contributions of 5.16, which was smaller than the bias of 9.15 for the linear model (Table 7). Therefore, the compound was predicted to be inactive. By contrast, the visualization of the correct prediction by the Tanimoto model (Figure 4 bottom) provides a different picture. All of the features made contributions of small magnitudes below 10% to the overall sum, and most of these contributions were positive. The feature mapping also showed that there was no single atom or bond in the molecule with a negative cumulative weight. In this case, the sum of feature contributions (4.46) clearly exceeded the relatively low bias of the Tanimoto model (3.46; Table 7), leading to a correct prediction of activity.

Figure 5 shows the results for an inactive test compound from the adenosine A2a receptor modeling, which was predicted to be active by the linear model. The visualization for the linear model (Figure 5 top) also identified the six-membered ring as a single dominant positive feature, although in this case it only accounted for less than 30% of the overall sum. Furthermore, there were a number of additional features with relatively small positive or negative contributions between 10% and 20%. Feature mapping identified a number of features with negative contributions centered at nitrogen and oxygen atoms throughout the molecule, similar to observations made for the linear model in Figure 4. However, in this case, the contribution of the six-membered ring and a part of the adjacent thiophene ring involving the sulfur atom clearly dominated the prediction, leading to a sum of 12.94 and a false positive assignment. The Tanimoto model correctly predicted this compound to be inactive. The visualization of this prediction (Figure 5 bottom) reveals the presence of many positive and negative contributions, especially from patterns. However, these contributions are of relatively small magnitude. As a result of the presence of a variety of positive and negative feature contributions in the Tanimoto model, which partly compensated for each other, the sum of contributions (3.15) for the compound in Figure 5 did not reach the model bias (3.46), leading to a correct prediction of inactivity. Furthermore, feature mapping showed that there were only few oxygen atoms with an overall negative weight.

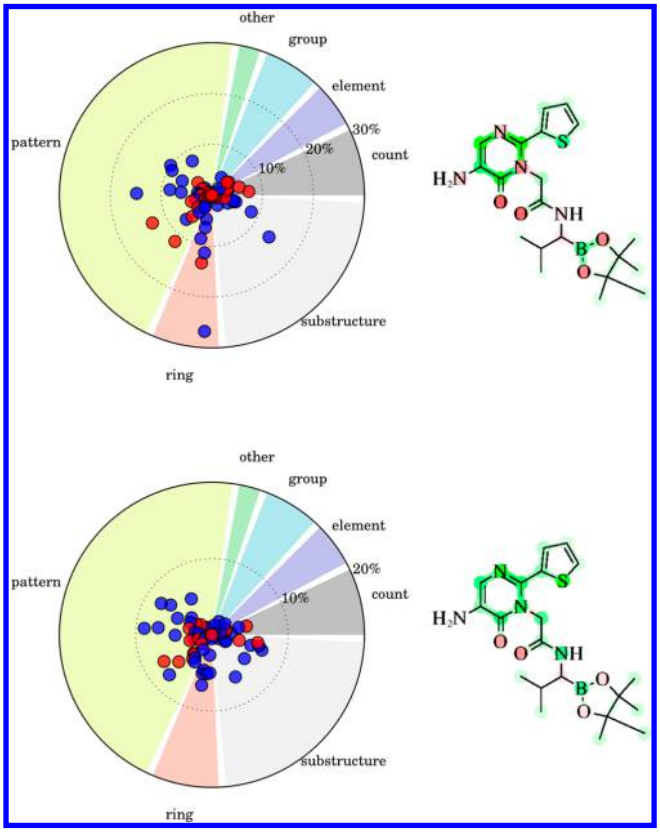


Figure 5. Visualization of predictions and feature mapping for ChEMBL compound 1113, represented according to Figure 4. This compound is a negative test example for prediction of adenosine A2a receptor ligands. Predictions: active (linear model) and inactive (Tanimoto model).

In general, features shared by predictions using the linear and Tanimoto kernels often made contributions of lesser magnitude to the Tanimoto model. This difference is a direct consequence of the denominator in the Tanimoto kernel, which weights the numbers of fingerprint bits set on for two compounds by the total number of possible commonly set bits. Thus, different from the linear kernel, the Tanimoto kernel further differentiates between compound pairs sharing a large number of bits but differing in their sizes and total numbers of bits set to 1.

Cannabinoid CB2 Receptor Ligands. Figure 6 shows an active cannabinoid CB2 receptor ligand that was predicted to be inactive by the linear model and active by the Tanimoto model. The visualization of the prediction using the linear model (Figure 6 top) highlights the dominance of a single feature (pattern "QSQ", where Q refers to any heteroatom) that makes a strong negative contribution of over 90% of the cumulative sum. In addition, a comparable number of other features with intermediate positive or negative contributions became apparent. Feature mapping revealed that the "QSQ" pattern was centered on an individual sulfur atom contained in this compound. However, the neighboring carbon and oxygen atoms made only minor or no negative contributions to the prediction, while the sulfur atom itself had a strong negative net effect. Furthermore, the nitrogen atom made a significant contribution to the negative prediction because it was a part of several features with moderately negative effects. This example illustrates the utility of the feature mapping to highlight focused contributions in cases where individual atoms participate in multiple features influencing a prediction. The prediction of this compound using the Tanimoto kernel is visualized at the bottom of Figure 6. In this case, many features with positive contributions of small magnitude were observed, but only few with negative contributions approaching the 10% limit, thus rationalizing the positive prediction. This was also consistent with the view obtained from feature mapping, which revealed that all of the atoms, including the nitrogen and sulfur, and all of the bonds had positive cumulative weights.

Figure 7 visualizes predictions for a negative test example subjected to cannabinoid CB2 receptor ligand modeling, which was predicted to be active by the linear model and inactive by the Tanimoto model. In this case, the prediction visualization using the linear kernel (Figure 7 top) shows different

DOI: 10.1021/acs.jcim.5b00175
J. Chem. Inf. Model. 2015, 55, 1136−1147

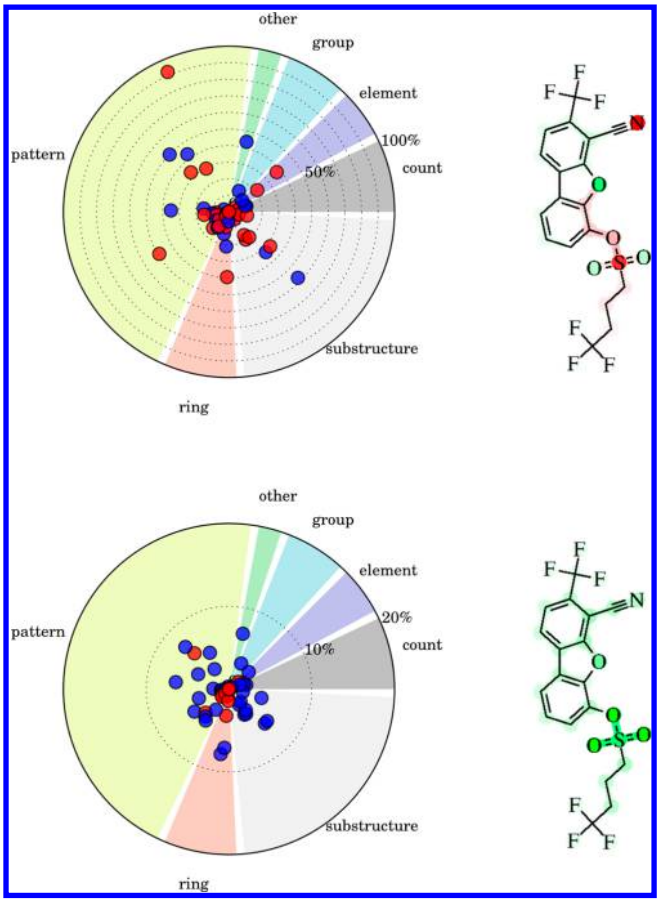


Figure 6. Visualization of predictions and feature mapping for ChEMBL compound 1834525, represented according to Figure 4. This compound is a cannabinoid CB2 receptor ligand. Predictions: inactive (linear), active (Tanimoto).

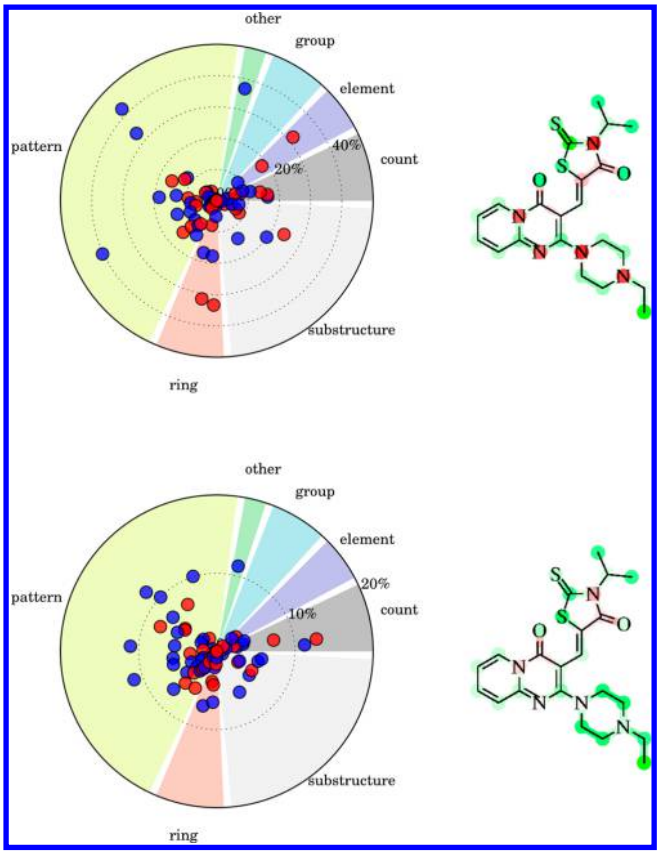


Figure 7. Visualization of predictions and feature mapping for ChEMBL compound 585041, represented according to Figure 4. This compound is a negative test example for prediction of cannabinoid CB2 receptor ligands. Predictions: active (linear), inactive (Tanimoto).

characteristics than the one discussed above. There were four features with strong positive contributions, three of which were patterns ("S=A", "A!N$A", and "NAO", where A refers to any aliphatic atom) and one was from the "other" group ("aromatic"). In addition, four features with negative contributions exceeding 30% of the overall sum included two rings ("N heterocycle", "ring"), a substructure ("OC(N)C"), and an element ("N"). Because two of the positive features and three of the negative features contained a nitrogen feature, mapping was crucial in this case to determine the contributions of these atoms. Mapping revealed that all of the nitrogen atoms ultimately made net contributions to the prediction of inactivity, whereas oxygen and sulfur atoms made contributions to the false prediction of activity, exceeding the bias of the linear model (Table 7). Interestingly, the feature mapping for the Tanimoto model (Figure 7 bottom) indicated that most of the atoms and patterns, except nitrogens and a few bond patterns, made positive contributions. This was also reflected in the prediction visualization, where the occurrence of a single nitrogen did not make any notable contribution; the only feature with a negative contribution exceeding the 10% limit was "A$A!O>1", which covered all of the nitrogen atoms in the compound. Hence, some of the nitrogens had a small cumulative negative weight, but overall there were many compensatory effects, and the sum of all contributions (2.15) was slightly smaller than the bias of the Tanimoto model (2.81; Table 7), leading to the correct prediction of inactivity. However, the numerical difference between the sum arising from multiple positive and negative feature contributions and the bias was relatively small in this case, consistent with the visualization in Figure 7.

Dopamine D2 Receptor Ligands. The active compound depicted in Figure 8 is small, consisting of only a condensed three-ring system. It was predicted to be inactive by the linear model and active by the Tanimoto model. The visualization of the prediction using the linear kernel (Figure 8 top) reveals a substructure feature ("CN(C)C") with a large positive contribution of more than 200% of the final sum. However, two patterns ("AN(A)A" and "AQ(A)A", where A refers to an aliphatic atom and Q to a heteroatom) with negative contributions of about 100% of the final sum nearly nullified this effect because the substructure CN(C)C matched both of these patterns. Aside from these features, the "aromatic" feature made the largest contribution. Feature mapping showed that cumulative negative contributions were mostly centered on the three nitrogen atoms, while positive contributions were distributed over the ring atoms (resulting in small atom-centric contributions). Overall, the negative and positive feature contributions were nearly compensatory, resulting in a cumulative feature contribution of 1.51, which was much smaller than the model bias of 5.90 (Table 7). The prediction using the Tanimoto model (Figure 8 bottom) was numerically vulnerable to boundary effects since the cumulative positive contribution of 2.89 only slightly exceeded the model bias of 2.83 (Table 7). However, the prediction visualization reveals a

DOI: 10.1021/acs.jcim.5b00175
J. Chem. Inf. Model. 2015, 55, 1136−1147

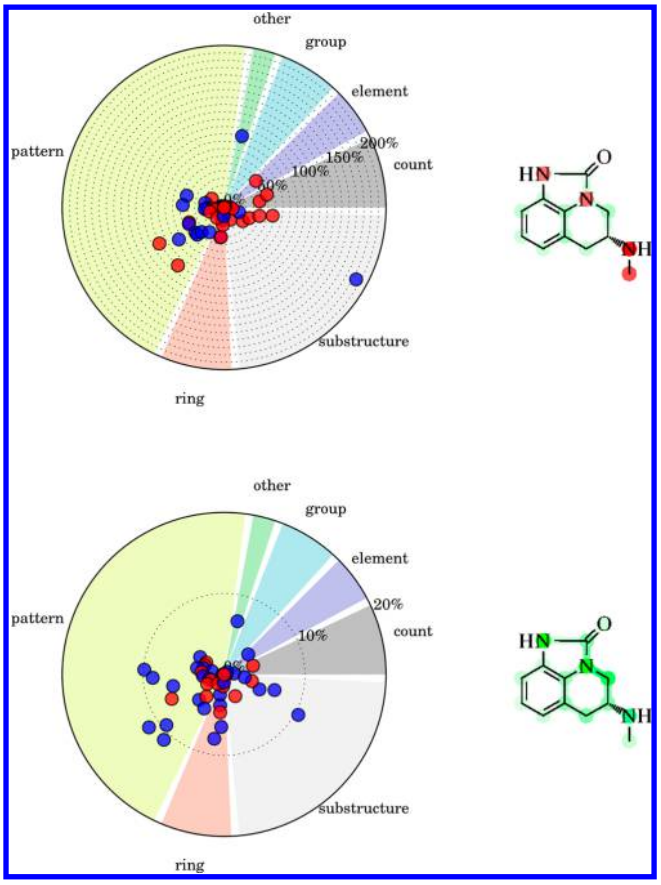


Figure 8. Visualization of predictions and feature mapping for ChEMBL compound 419792, represented according to Figure 4. This compound is a dopamine D2 receptor ligand. Predictions: inactive (linear), active (Tanimoto).

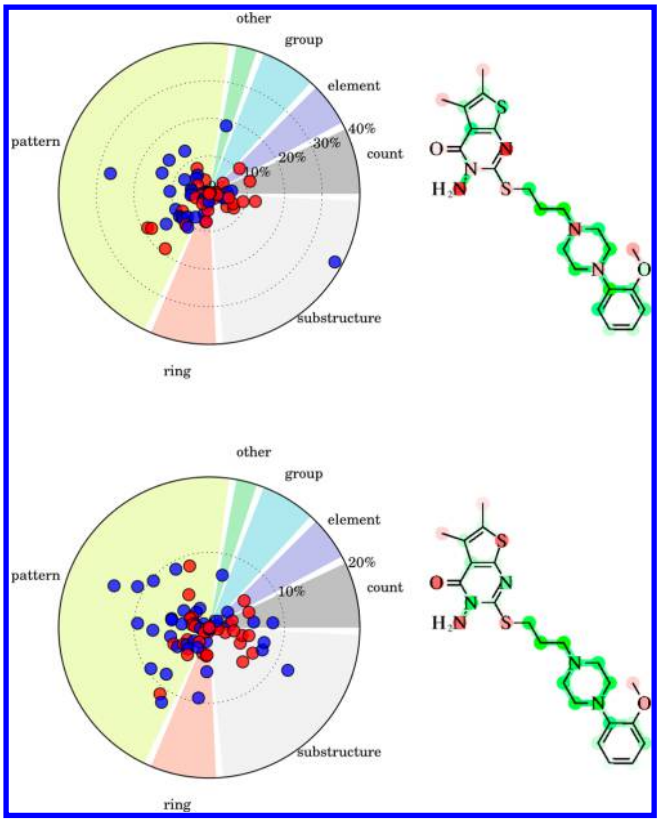


Figure 9. Visualization of predictions and feature mapping for ChEMBL compound 12028, represented according to Figure 4. This compound is a negative test example for prediction of dopamine D2 receptor ligands. Predictions: active (linear), inactive (Tanimoto).

variety of positive feature contributions (more so than negative ones), also including the "CN(C)C" substructure at the 10% level. However, all of the contributions were of low magnitude (with a maximum value of 0.33, corresponding to 11.34% of the overall sum). Nonetheless, despite the overall low cumulative feature contribution comparable to the model bias, feature mapping revealed only positive net contributions at atoms and bonds across the molecule.

Figure 9 visualizes predictions for a negative test example in dopamine D2 receptor ligand modeling that was predicted to be active by the linear model and inactive by the Tanimoto model. This molecule is much larger than the active compound in Figure 8. The visualization for the linear model (Figure 9 top) reveals the largest positive contributions to come from a substructure ("CN(C)C"), a pattern ("QN"), and another feature ("aromatic"), while negative contributions mostly originated from patterns ("AQ(A)A", "AN(A)A", and "QQ"). Again, contributions of these features largely compensated for each other because the substructure CN(C)C also matched AQ(A)A and AN(A)A and the same fragment matched patterns QN and QQ. Nonetheless, the sum of contributions from predictions using the linear kernel in Figure 9 was 8.14, which clearly exceeded the model bias of 5.90 (Table 7), leading to a false positive prediction. Feature mapping showed that negative contributions, mostly from nitrogen atoms, were too small to match cumulative positive contributions from the linker and ring systems of the compound. The visualization of the prediction using the Tanimoto kernel (Figure 9 bottom) provides a different picture. In this case, many features with positive or negative contributions of varying magnitudes were identified, and there were no individual features that largely determined the predictions. Feature mapping revealed the presence of multiple positive and negative contributions in corresponding regions of the negative test compound using the linear and Tanimoto kernels. However, in the case of the Tanimoto model, nitrogen atoms mostly made positive contributions, while oxygen and sulfur atoms made negative contributions, which was different from the linear model. Furthermore, the contributions of the aromatic ring systems to the prediction using the Tanimoto kernel were of lesser magnitude compared with the linear kernel. Overall, the Tanimoto kernel yielded a number of compensatory positive and negative feature contributions, and the final sum (2.12) did not reach the model bias of 2.83 (Table 7).

MAP Kinase p38 Alpha Inhibitors. Figure 10 shows an active MAP kinase p38 alpha inhibitor that was incorrectly predicted to be inactive by the linear kernel and correctly predicted to be active by the Tanimoto kernel. The prediction visualization revealed that there were many features in the linear prediction that contributed in a similar way. The strongest positive features were the substructure "NH", the ring feature "aromatic ring >1", and the pattern "NA(A)A". The features with strongest negative contributions included the substructure "C−N" and the pattern "NACH2A". Overall, the mapping showed that nitrogen and oxygen atoms made overall positive contributions, while the linker in the lower part of the molecule made the only considerable negative contribution. The overall sum of feature contributions was 2.08, which was

DOI: 10.1021/acs.jcim.5b00175
J. Chem. Inf. Model. 2015, 55, 1136−1147

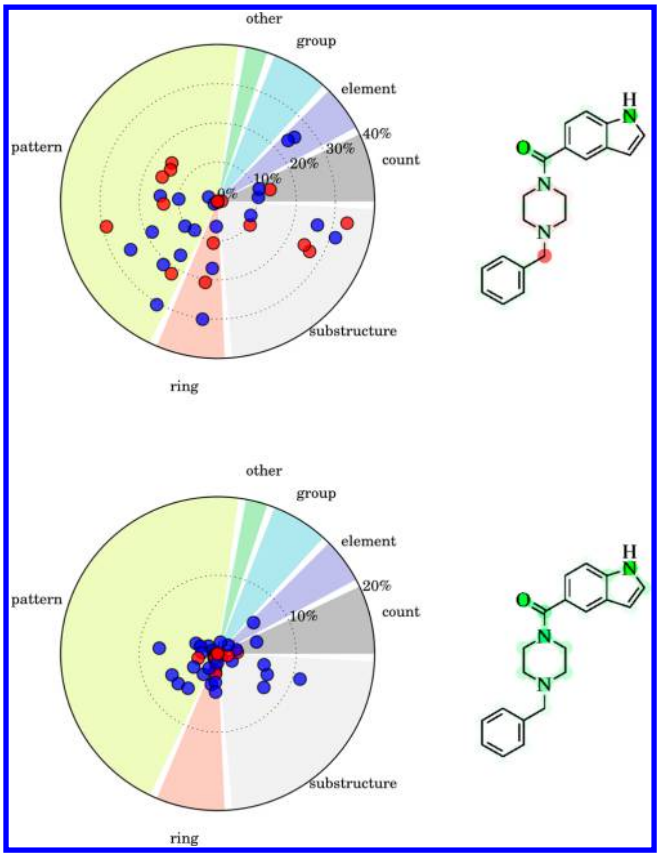


Figure 10. Visualization of predictions and feature mapping for ChEMBL compound 320069, represented according to Figure 4. This compound is a positive test example for prediction of MAP kinase p38 alpha ligands. Predictions: inactive (linear), active (Tanimoto).

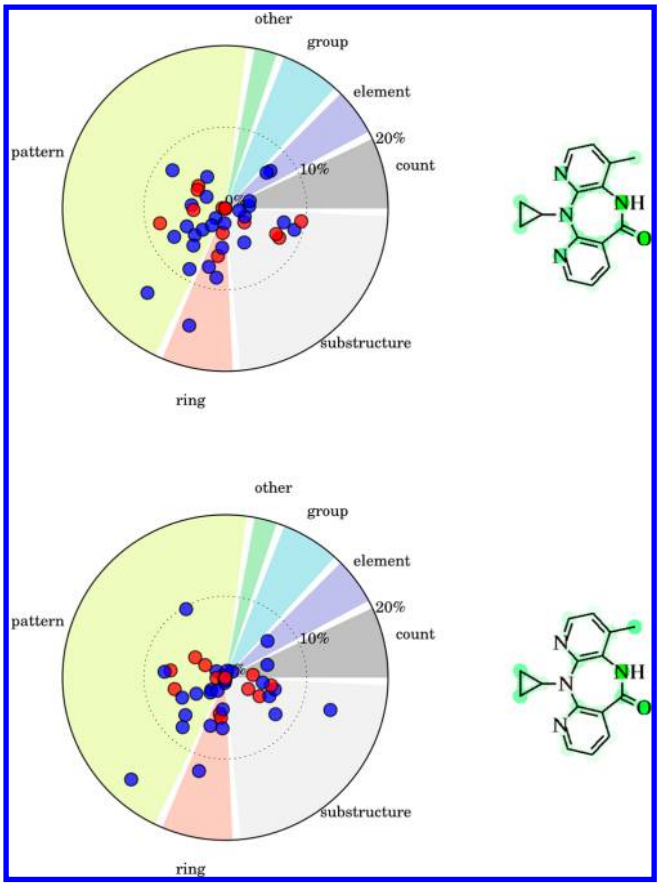


Figure 11. Visualization of predictions and feature mapping for ChEMBL compound 57, represented according to Figure 4. This compound is a negative test example for prediction of MAP kinase p38 alpha ligands. Predictions: active (linear), inactive (Tanimoto).

considerably smaller than the threshold of 5.74, meaning that the few cumulative positive contributions were not sufficient to yield a prediction of activity. By contrast, the visualization of the Tanimoto prediction revealed a different picture. Here only one substructure ("NH") having a considerable positive contribution was identified, in addition to many smaller positive contributions. There were very few if any negative contributions, as was also confirmed by feature mapping, which identified only cumulative positive contributions. Here the sum of the feature contributions of 4.33 exceeded the model bias by 1 (Table 7).

Furthermore, Figure 11 shows the prediction visualization and mappings for an inactive compound that was predicted to be active by the linear model and inactive by the Tanimoto model. The linear prediction visualization identified two important positive contributions, including a seven-membered ring and the pattern "A!:A:A!:A" (where ":" refers to an aromatic bond and "!" denotes negation). Many other features contributed only less than 10% to the overall sum. Here the prediction visualization showed that there were more features with positive contributions than negative ones. Features covering nearly the entire compound contributed to the positive prediction, leading to a large sum of 7.35 (Table 7). By contrast, the Tanimoto kernel's contributions yielded a sum of 2.84, which did not reach the model bias of 3.33, leading to a correct inactive prediction. In this case, three features contributed more than 10% to the overall sum: the pattern "A!:A:A!:A", the three-membered ring, and the substructure "NH". There were also several features with small positive or negative contributions. Feature mapping revealed that three nitrogen atoms made much smaller contributions compared with the linear case, leading to a smaller sum and thus the prediction of inactivity.

Taken together, the visualizations of the exemplary predictions in Figures 4−11 in combination with feature mapping helped to rationalize kernel-dependent differences in activity predictions. While several predictions using the simple linear kernel were dominated by individual feature contributions, the Tanimoto kernel — given its design, as discussed above — better differentiated feature contributions and their relative magnitudes. The Tanimoto kernel also generally reduced the magnitude of feature contributions, thereby balancing the influence of individual features on the predictions. Feature mapping complemented the visualization of predictions by focusing on atoms, bonds, or other substructures (e.g., rings) that were involved in multiple features and accounting for net effects. The exemplary predictions summarized in Table 7 also illustrate that cumulative feature contributions were rarely negative, even for compounds correctly predicted to be inactive — a previously unobserved effect. In these cases, the positive cumulative contributions were smaller than the model biases.

Variation of Inactive Training Compounds. All of the calculations discussed herein with negative training examples taken from ChEMBL were repeated with an equally sized negative training set randomly selected from ZINC. In order to analyze the effect of different inactive training sets on our visualization method, prediction visualization and feature

DOI: 10.1021/acs.jcim.5b00175
J. Chem. Inf. Model. 2015, 55, 1136−1147

mappings of active compounds were compared. Figure 12 shows the comparison of the linear and Tanimoto models for

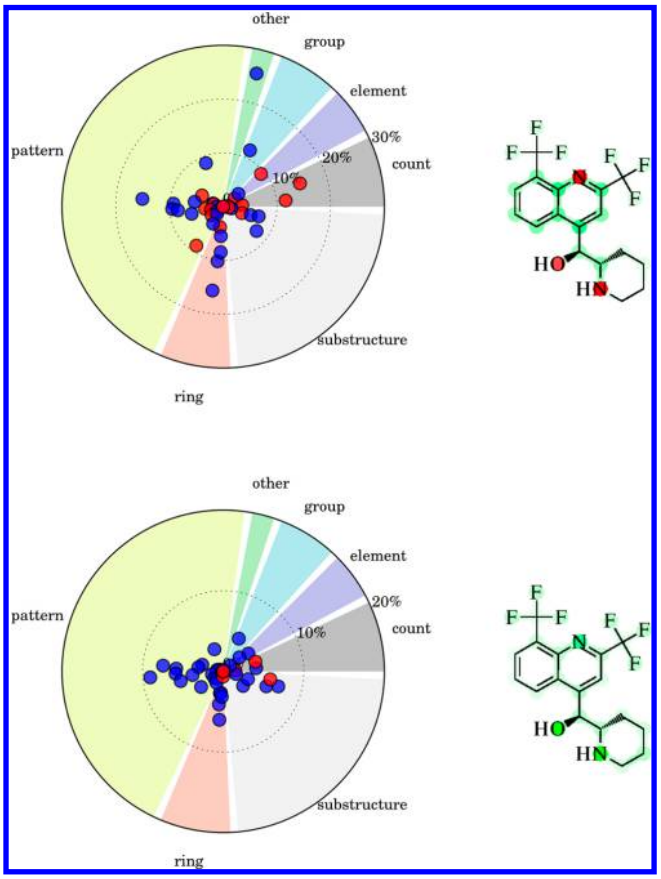


Figure 12. Visualization of predictions and feature mapping for the adenosine A2a receptor ligand from Figure 4 obtained using a subset of ZINC compounds as inactive training instances and the (top) linear and (bottom) Tanimoto models.

the adenosine A2a receptor ligand analyzed in Figure 4 using the ZINC training set. The compound was predicted to be inactive by the linear model and active by the Tanimoto model, regardless of the choice of the inactive training set. Feature mapping revealed that the same atoms and bonds contributed positively or negatively to the prediction in both cases. However, the prediction visualization showed that the features leading to these cumulative contributions differed. For instance, the most important positive and negative features for the linear ChEMBL-based models were the six-membered ring and the N heterocycle, respectively, as discussed above. By contrast, for the linear ZINC-based model, the "aromatic" feature from the "other" group was the most important positive feature and the "N>1" feature from the "count" group the most important negative feature. However, the Tanimoto models derived using ChEMBL and ZINC subsets did not differ notably.

Figure 13 shows the active cannabinoid CB2 receptor ligand from Figure 6 together with its prediction visualization and feature mappings for the linear and Tanimoto models based upon the inactive training set from ZINC. The compound was predicted to be inactive by the linear model and active by the Tanimoto model. However, both prediction visualization and feature mapping of the linear model differed from those shown in Figure 6. For example, the pattern "S=A" made by far the largest positive contribution in the linear ZINC model. This

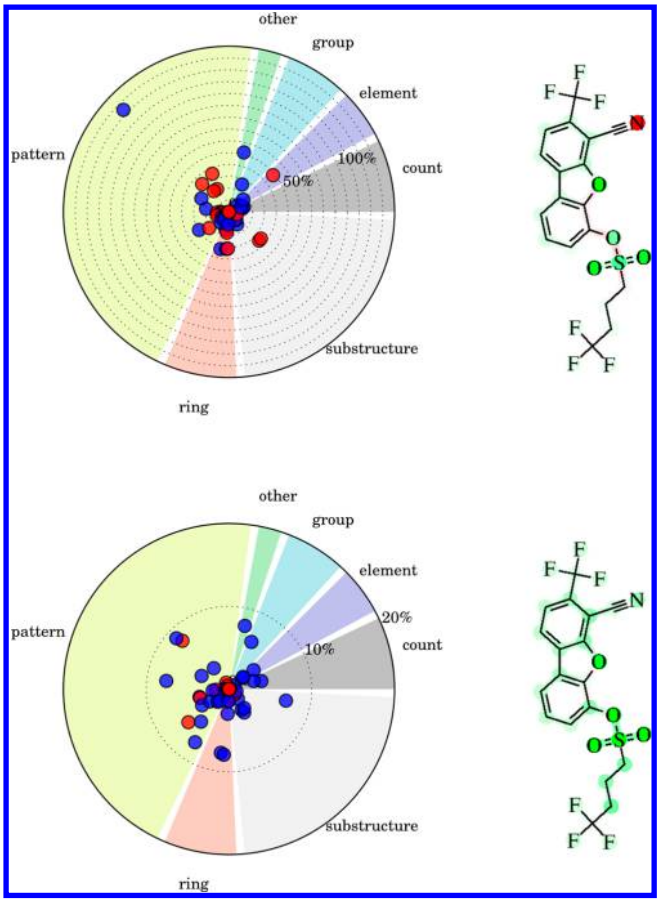


Figure 13. Visualization of predictions and feature mapping for the cannabinoid CB2 receptor ligand from Figure 6 obtained using a subset of ZINC compounds as inactive training instances and the (top) linear and (bottom) Tanimoto models.

contribution caused a change in the feature mapping from the sulfur atom making a negative contribution (ChEMBL-based model, Figure 6) to a positive contribution (ZINC-based model, Figure 13). Other features with contributions of ~50% to the overall sum in the ChEMBL model made smaller contributions to the ZINC model. By contrast, the Tanimoto models using the ChEMBL and ZINC subsets displayed very similar characteristics in prediction visualization and feature mapping. The examples in Figures 12 and 13 show how the choice of negative training data might influence an SVM model and how prediction visualization and feature mapping can be used to analyze this influence.

CONCLUSIONS

In this study, we have introduced a visualization method for SVM predictions using the Tanimoto kernel compared with the linear kernel. Our analysis has revealed how activity predictions are determined by contributions from varying numbers of fingerprint features. The study can be extended to other kernel functions for which feature contributions can be expressed as sums. Visualization complemented by feature mapping provides a direct diagnostic for SVM models and reduces the black box character of SVM predictions. An implementation of the visualization approach introduced herein has been made freely available to aid in the assessment of SVM models and their successes and failures.

DOI: 10.1021/acs.jcim.5b00175
J. Chem. Inf. Model. 2015, 55, 1136−1147

AUTHOR INFORMATION

Corresponding Author

*Tel: +49-228-2699-306. Fax: +49-228-2699-341. E-mail: bajorath@bit.uni-bonn.de.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The use of OpenEye's OEChem and OEDepict Toolkit was made possible by their free academic licensing program.

REFERENCES

(1) Heikamp, K.; Bajorath, J. Support Vector Machines for Drug Discovery. *Expert Opin. Drug Discovery* **2014**, *9*, 93−104.

(2) Geppert, H.; Humrich, J.; Stumpfe, D.; Gärtner, T.; Bajorath, J. Ligand Prediction from Protein Sequence and Small Molecule Information Using Support Vector Machines and Fingerprint Descriptors. *J. Chem. Inf. Model.* **2009**, *49*, 767−779.

(3) Erhan, D.; L'Heureux, P.-J.; Yue, S. Y.; Bengio, Y. Collaborative Filtering on a Family of Biological Targets. *J. Chem. Inf. Model.* **2006**, *46*, 626−635.

(4) Jacob, L.; Vert, J.-P. Protein−Ligand Interaction Prediction: An Improved Chemogenomics Approach. *Bioinformatics* **2008**, *24*, 2149−2156.

(5) Kawai, K.; Fujishima, S.; Takahashi, Y. Predictive Activity Profiling of Drugs by Topological-Fragment-Spectra-Based Support Vector Machines. *J. Chem. Inf. Model.* **2008**, *48*, 1152−1160.

(6) Heikamp, K.; Bajorath, J. Prediction of Compounds with Closely Related Activity Profiles Using Weighted Support Vector Machine Linear Combinations. *J. Chem. Inf. Model.* **2013**, *53*, 791−801.

(7) Wassermann, A. M.; Heikamp, K.; Bajorath, J. Potency-Directed Similarity Searching Using Support Vector Machines. *Chem. Biol. Drug Des.* **2011**, *77*, 30−38.

(8) Rathke, F.; Hansen, K.; Brefeld, U.; Müller, K.-R. StructRank: A New Approach for Ligand-Based Virtual Screening. *J. Chem. Inf. Model.* **2011**, *51*, 83−92.

(9) Ralaivola, L.; Swamidass, S. J.; Saigo, H.; Baldi, P. Graph Kernels for Chemical Informatics. *Neural Networks* **2005**, *18*, 1093−1110.

(10) Rogers, D. J.; Tanimoto, T. T. A Computer Program for Classifying Plants. *Science* **1960**, *132*, 1115−1118.

(11) Navia-Vázquez, A.; Parrado-Hernández, E. Support Vector Machine Interpretation. *Neurocomputing* **2006**, *69*, 1754−1759.

(12) Rosenbaum, L.; Hinselmann, G.; Jahn, A.; Zell, A. Interpreting Linear Support Vector Machine Models with Heat Map Molecule Coloring. *J. Cheminf.* **2011**, *3*, No. 11.

(13) Franke, L.; Byvatov, E.; Werz, O.; Steinhilber, D.; Schneider, P.; Schneider, G. Extraction and Visualization of Potential Pharmacophore Points Using Support Vector Machines: Application to Ligand-Based Virtual Screening for COX-2 Inhibitors. *J. Med. Chem.* **2005**, *48*, 6997−7004.

(14) Devos, O.; Ruckebusch, C.; Durand, A.; Duponchel, L.; Huvenne, J.-P. Support Vector Machines (SVM) in Near Infrared (NIR) Spectroscopy: Focus on Parameters Optimization and Model Interpretation. *Chemom. Intell. Lab. Syst.* **2009**, *96*, 27−33.

(15) Mohr, J.; Jain, B.; Sutter, A.; Laak, A. T.; Steger-Hartmann, T.; Heinrich, N.; Obermayer, K. A Maximum Common Subgraph Kernel Method for Predicting the Chromosome Aberration Test. *J. Chem. Inf. Model.* **2010**, *50*, 1821−1838.

(16) Carlsson, L.; Helgee, E. A.; Boyer, S. Interpretation of Nonlinear QSAR Models Applied to Ames Mutagenicity Data. *J. Chem. Inf. Model.* **2009**, *49*, 2551−2558.

(17) Martens, D.; Huysmans, J.; Setiono, R.; Vanthienen, J.; Baesens, B. Rule Extraction from Support Vector Machines: An Overview of Issues and Application in Credit Scoring. *Stud. Comput. Intell.* **2008**, *80*, 33−63.

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

(19) Baehrens, D.; Schroeter, T.; Harmeling, S.; Kawanabe, M.; Hansen, K.; Müller, K.-R. How To Explain Individual Classification Decisions. *J. Mach. Learn. Res.* **2010**, *11*, 1803−1831.

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

(21) Cortes, C.; Vapnik, V. N. Support Vector Networks. *Mach. Learn.* **1995**, *20*, 273−297.

(22) Kuhn, H. W.; Tucker, A. W. Nonlinear Programming. *Proc. Berkeley Symp. Math., Stat. Probab., 2nd* **1950**, 481−492.

(23) Boser, B. E.; Guyon, I. M.; Vapnik, V. N. A Training Algorithm for Optimal Margin Classifiers. *Proc. Annu. Workshop Comput. Learn. Theory, 5th* **1992**, 144−152.

(24) Mercer, J. Functions of Positive and Negative Type, and their Connection with the Theory of Integral Equations. *Philos. Trans. R. Soc. London, Ser. A* **1909**, *209*, 415−446.

(25) Gärtner, T.; Flach, P.; Wrobel, S. On Graph Kernels: Hardness Results and Efficient Alternatives. In *Learning Theory and Kernel Machines*; Springer: Berlin, 2003.

(26) Kashima, H.; Tsuda, K.; Inokuchi, A. Marginalized Kernels between Labeled Graphs. *Proc. Int. Conf. Mach. Learn., 20th* **2003**, 321−328.

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

(28) Bento, A. P.; Gaulton, A.; Hersey, A.; Bellis, L. J.; Chambers, J.; Davies, M.; Kruger, F. A.; Light, Y.; Mak, L.; McGlinchey, S.; Nowotka, M.; Papadatos, G.; Santos, R.; Overington, J. P. The ChEMBL Bioactivity Database: An Update. *Nucleic Acids Res.* **2014**, *42*, 1083−1090.

(29) 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.

(30) Irwin, J. J.; Sterling, T.; Mysinger, M. M.; Bolstad, E. S.; Coleman, R. G. ZINC: A Free Tool To Discover Chemistry for Biology. *J. Chem. Inf. Model.* **2012**, *52*, 1757−1768.

(31) *MACCS Structural Keys*; Accelrys: San Diego, CA, 2011.

(32) *OEChem Toolkit*, version 2.0.2.; OpenEye Scientific Software: Santa Fe, NM; http://www.eyesopen.com.

(33) RDKit: Open-Source Cheminformatics. http://www.rdkit.org.

(34) Morik, K.; Brockhausen, P.; Joachims, T. Combining Statistical Learning with a Knowledge-Based Approach—A Case Study in Intensive Care Monitoring. *Proc. Int. Conf. Mach. Learn., 16th* **1999**, 268−277.

(35) Hsu, C.-W; Chang, C.-C.; Lin, C.-J. *A Practical Guide to Support Vector Classification*. Technical Report; Department of Computer Science, National Taiwan University: Taipei, Taiwan, 2003.

(36) Joachims, T. Making Large-Scale Support Vector Machine Learning Practical. In *Advances in Kernel Methods*; Schölkopf, B., Burges, C. J. C., Smola, A. J., Eds.; MIT Press: Cambridge, MA, 1999; pp 169−184.

(37) *OEDepict Toolkit*, version 2.2.4.; *OpenEye Scientific Software*: Santa Fe, NM; http://www.eyesopen.com.

(38) Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **2007**, *9*, 90−95.

(39) Balfer, J.; Bajorath, J. Visualization and Interpretation of Support Vector Machine Activity Predictions. DOI: 10.5281/zenodo.17718.

(40) 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.

(41) Polishchuk, P. G.; Kuz'min, V. E.; Artemenko, A. G.; Muratov, E. N. Universal Approach for Structural Interpretation of QSAR/QSPR Models. *Mol. Inf.* **2013**, *32*, 843−853.

DOI: 10.1021/acs.jcim.5b00175
J. Chem. Inf. Model. 2015, 55, 1136−1147