Journal of Chemical Information and Modeling

Article

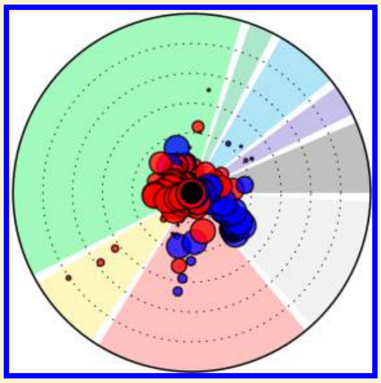
pubs.acs.org/jcim

Introduction of a Methodology for Visualization and Graphical Interpretation of Bayesian Classification Models

Jenny Balfer and Jürgen Bajorath*

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

ABSTRACT: Supervised machine learning models are widely used in chemoinformatics, especially for the prediction of new active compounds or targets of known actives. Bayesian classification methods are among the most popular machine learning approaches for the prediction of activity from chemical structure. Much work has focused on predicting structure–activity relationships (SARs) on the basis of experimental training data. By contrast, only a few efforts have thus far been made to rationalize the performance of Bayesian or other supervised machine learning models and better understand why they might succeed or fail. In this study, we introduce an intuitive approach for the visualization and graphical interpretation of naïve Bayesian classification models. Parameters derived during supervised learning are visualized and interactively analyzed to gain insights into model performance and identify features that determine predictions. The methodology is introduced in detail and applied to assess Bayesian modeling efforts and predictions on compound data sets of varying structural complexity. Different classification models and features determining their performance are characterized in detail. A prototypic implementation of the approach is provided.



INTRODUCTION

Machine learning models are used for a variety of applications in chemoinformatics including, for instance, the prediction of compound activity and other molecular properties or biological targets of known actives.[1−3] When a set of known positive (active) and negative (inactive) training compounds is available, supervised machine learning is an approach of choice for building predictive models of activity.[1,2] In the chemoinformatics community, the currently most frequently applied supervised machine learning methods include random forests, support vector machines, and Bayesian classifiers.[1,2] Random forest models utilize ensembles of decision trees to arrive at consensus predictions, support vector machines derive separating hyperplanes for class label prediction in feature spaces of increasing dimensionality, and Bayesian classifiers are probabilistic models based upon Bayes theorem.

In general, supervised learning is used to build complex models from training data that go beyond the derivation of simple rules to determine a classification outcome. Model building requires the definition of a suitable molecular descriptor space for classification and the selection of preferred models on the basis of preset performance criteria. If sufficient training data is available and meaningful reference (feature) spaces can be generated, effective models can often be derived for a variety of classification or regression tasks.[4−7]

Areas in which supervised learning is applied can roughly be divided into those in which a computer should learn a concept that is intuitively known to humans as opposed to those where the concept itself is not fully understood by human experts. An example for the first area is image classification. Humans can usually identify and distinguish different objects in images. Yet, successful recognition is the result of very complex reasoning and neural functions, which cannot be easily transferred to a computer.[8] Hence, supervised classification is applied in such situations to let the computer "learn" the concept from examples. However, many chemoinformatics problems fall into the second area mentioned above. For example, even an experienced medicinal chemist can typically not predict the activity of given compounds against biological targets in a consistent manner and without error. Thus, much research is dedicated to rationalizing and predicting structure−activity relationships (SAR)[9−11] as there are no generally applicable rules governing compound-target interactions that could be consistently applied. Even in the presence of significant amounts of experimental data, the exact mechanism of compound-target interactions is often difficult to determine.[12] Therefore, if a machine learning model for activity prediction can be derived, it should be important to understand the characteristics of the model that determine its decisions. However, this is in general difficult to accomplish. Clearly, obtaining such insights would help to reduce or eliminate the well-known "black box" character of many machine learning approaches, which often limits their utility. In interdisciplinary research, gaining insights into the mechanisms by which computer models function is often a prerequisite of their acceptance and for the willingness to build experimental projects around predictions. Hence, the importance of chemical interpretability of machine learning models and their predictions should not be underestimated.

Received: July 9, 2014
Published: August 19, 2014

Molecular feature spaces used in chemoinformatics are typically large and high-dimensional, as millions of biologically relevant compounds and thousands of chemical descriptors are available.[13] For the navigation of such feature spaces and property prediction, naïve Bayesian classifiers are often applied.[14−25] Their popularity can be attributed to their relatively simplistic design, the ability to efficiently operate on large and high-dimensional data sets, and their limited sensitivity to data noise; an important aspect for chemoinformatics applications.[14,15] In recent years, naïve Bayesian classifiers have been applied to identify therapeutically relevant targets[16] as well as novel active compounds for given targets,[17−20] further improve docking scores,[15,21−23] or predict absorption, distribution, metabolism, and excretion (ADME) properties[24] and multidrug resistance reversal activity.[25] For such studies, compounds have mostly been represented using binary fingerprints.[14−20,26]

Although a number of successful naïve Bayesian models have been reported, only very few studies have thus far attempted to address the question how exactly these models work and why they might succeed or fail. As Klon et al. point out in their study to predict ADME properties, "understanding why a compound has undesirable ADME characteristics is just as important as knowing that it does".[24] These authors have also aimed to rationalize their classification models. For instance, it was attempted to explain the success of a naïve Bayesian classifier in enriching favorable docking scores by training alternative models on only subsets of features from preferred models.[21] Other investigators have addressed the interpretability issue by using intuitive molecular representations such as chemical fragment descriptors[19] or by focusing on specific compounds whose activity could only be predicted using naïve Bayesian classification or other machine learning approaches.[27]

In this study, we introduce an intuitive approach for the visualization and graphical interpretation of naïve Bayesian classification models. Previous work on graphical interpretation of machine learning models has primarily focused on depicting features in heat maps[28] or similarity maps[29] that are important for prediction of individual molecules. The methodology introduced herein also enables the assessment of individual predictions but goes far beyond the analysis of single compounds by providing a visualization scheme for an entire classification model. Furthermore, it also reveals the contributions of features that are absent in test compounds.

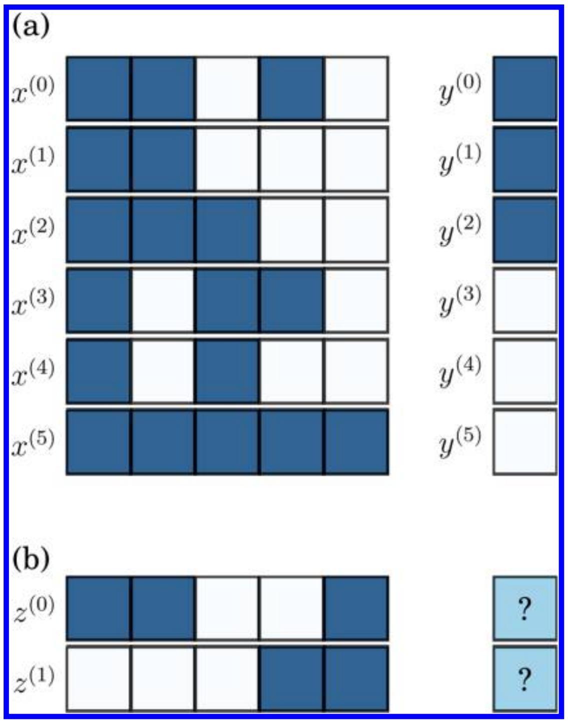


Figure 1. Motivating example. Shown is a theoretical "minimalist" example of a training and test set for supervised classification, represented as binary features. Blue squares indicate that a feature is set and white squares that it is not set. (a) Six training examples are given; three of which are positive, and the remaining three are negative. (b) The test set contains two examples with unknown class labels.

Table 1. Estimated Parameters for a Naïve Bayesian Model[a]

d	$P(x_d = 1\|y = 0)$	$P(x_d = 1\|y = 1)$	$P(y = 1)$
0	0.9688	0.9688	
1	0.3438	0.9688	
2	0.9688	0.3438	0.5
3	0.6562	0.3438	
4	0.3438	0.0313	

[a] Reported are the conditional feature probabilities and the prior probability for the model of the motivating example discussed in the text.

CONCEPTS AND METHODS

Naïve Bayesian classification. The naïve Bayesian classifier makes use of Bayes' theorem to predict the probability $P(y|x)$ of an instance x to belong to class y:[30]

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)} \tag{1}$$

It is called naïve because it assumes all features x_d in x to be independent of each other;[31] applying this feature independence assumption, eq 1 can be rewritten as

$$P(y|x) = \frac{\prod_d P(x_d|y)P(y)}{\prod_d P(x_d)} \tag{2}$$

To build a naïve Bayesian model, a training set of labeled instances with different class labels is utilized for supervised learning. Although there are no principal assumptions concerning the nature of x and y, one often focuses on binary features and class labels, i.e., $x,y \in \{0,1\}$, for example, "active" vs "inactive". In this case, given a set of n training instances X and corresponding labels Y, the terms required for naïve Bayesian classification can be estimated as follows:[30]

$$P(x_d = 1|y = \hat{y}) = \frac{\sum_i x_d^{(i)}\delta_{y^{(i)}\hat{y}} + \alpha}{\sum_i \delta_{y^{(i)}\hat{y}} + 2\alpha} \tag{3}$$

$$P(y = 1) = \frac{\sum_i y^{(i)}}{n} \tag{4}$$

Here, δ_{ij} is the Kronecker delta function, which is 1 for $i = j$ and 0 otherwise. The notations $x^{(i)}$ and $y^{(i)}$ refer to the i'th training instance and label, respectively. The term α is a Laplacian smoothing factor used to prevent the introduction of ill-defined probabilities, e.g., if a feature is never set in the training data. Since both class labels y and features x_d are binary, we can infer

$$P(x_d = 0|y = \hat{y}) = 1 - P(x_d = 1|y = \hat{y}) \tag{5}$$

$$P(y = 0) = 1 - P(y = 1) \tag{6}$$

Table 2. Odds Ratios for a Naïve Bayesian Model[a]

d	OR_d	$\log \mathrm{OR}_d$
0	1.00	0.00
1	2.82	1.04
2	0.35	−1.04
3	0.52	−0.65
4	0.09	−2.40

[a] Reported are the odds ratios of features x_0−x_4 in the model of the motivating example. The closer the odds ratio of a given feature is to 1, the smaller is its influence on the classification.

The meaning of these equations can be illustrated by a simple example. Let us derive a naïve Bayesian model from a theoretical training set of three positive and negative examples each, which are each represented by five features, as illustrated in Figure 1a. In the following, we refer to this example as the "motivating example" because it can be used to illustrate the opportunities of model visualization.

By applying eqs 3 and 4 and by setting $\alpha = 0.1$, the probabilities reported in Table 1 are obtained. Using eqs 5 and 6, all missing probabilities are derived. One now can make predictions for test

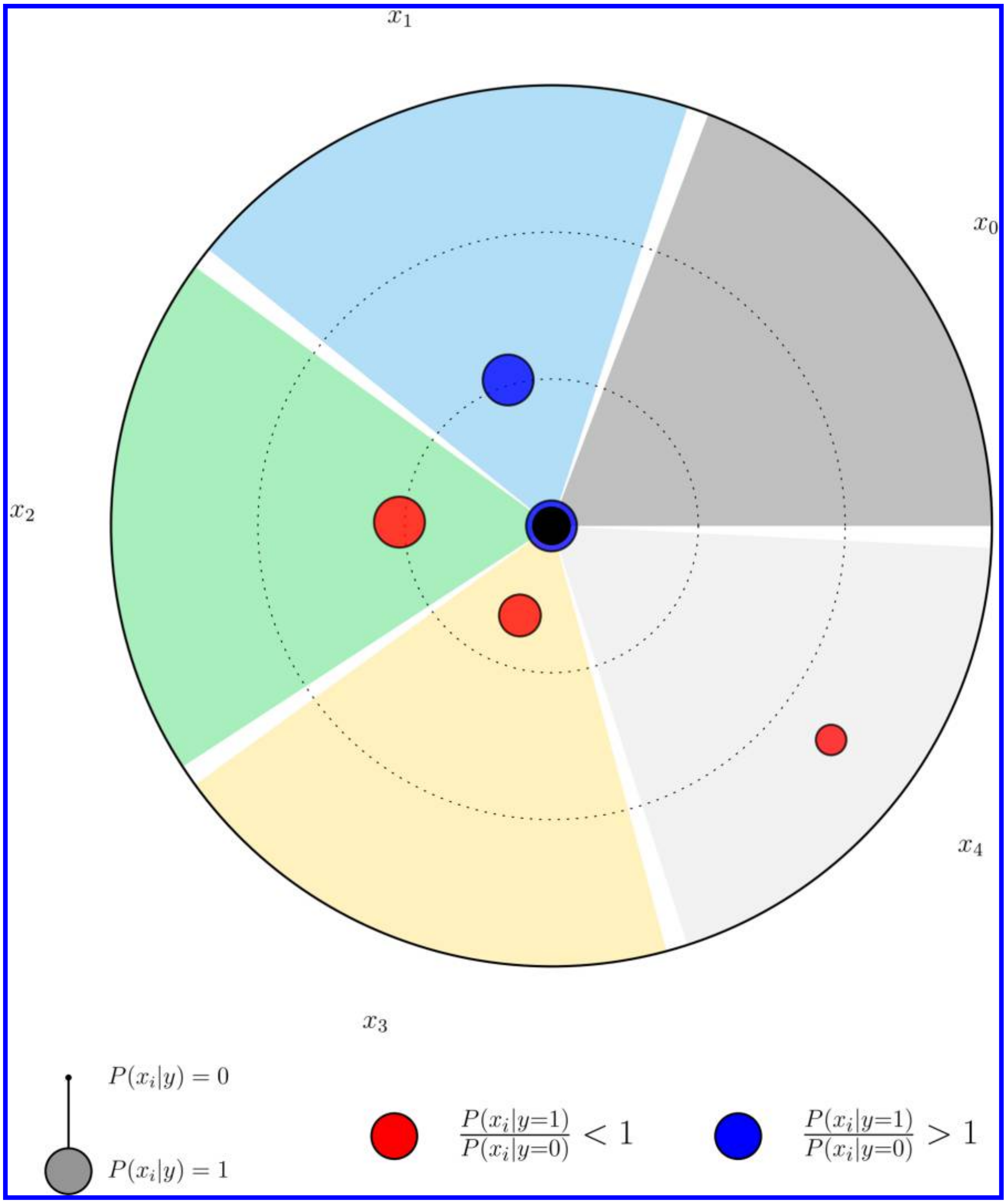


Figure 2. Principles of model visualization. The model for the motivating example is visualized. Each point represents a feature, and the distance to the pole corresponds to its absolute log odds ratio. Positive and negative influence on the classification is indicated by color coding and likelihood by size scaling, as detailed in the text.

instances such as $z^{(0)}$ and $z^{(1)}$ in Figure 1b. For each instance, we can calculate the class likelihood $P(z^{(i)}|y = \hat{y})$ and the evidence $P(z^{(i)})$ as follows:

$$P(z^{(0)}|y = 0) = \prod_d P(x_d = z_d^{(0)}|y = 0) = 0.00123$$

$$P(z^{(0)}|y = 1) = \prod_d P(x_d = z_d^{(0)}|y = 1) = 0.01263$$

$$P(z^{(1)}|y = 0) = \prod_d P(x_d = z_d^{(1)}|y = 0) = 0.00014$$

$$P(z^{(1)}|y = 1) = \prod_d P(x_d = z_d^{(1)}|y = 1) = 0.00001$$

$$P(z^{(0)}) = \sum_{\hat{y}} P(\hat{y}) \prod_d P(x_d = z_d^{(0)}|\hat{y}) = 0.0069$$

$$P(z^{(1)}) = \sum_{\hat{y}} P(\hat{y}) \prod_d P(x_d = z_d^{(1)}|\hat{y}) = 0.0001$$

The posterior probabilities are then given as

$$P(y = 0|z^{(0)}) = \frac{P(z^{(0)}|y = 0)P(y = 0)}{P(z^{(0)})} = 0.0887$$

$$P(y = 1|z^{(0)}) = \frac{P(z^{(0)}|y = 1)P(y = 1)}{P(z^{(0)})} = 0.9113$$

$$P(y = 0|z^{(1)}) = \frac{P(z^{(1)}|y = 0)P(y = 0)}{P(z^{(1)})} = 0.9545$$

$$P(y = 1|z^{(1)}) = \frac{P(z^{(1)}|y = 1)P(y = 1)}{P(z^{(1)})} = 0.0455$$

In this case, the positive class label would be predicted for the first and the negative class label for the second test instance.

Model Interpretation. The motivating example discussed above illustrates two important points: First, for classification, the marginal probability $P(z)$ is only used as a normalization factor and can hence be omitted if knowledge of exact posterior probabilities is not required. In fact, it has been shown that successful naïve Bayesian classification models often produce rather poor probability estimates.[32] Given that exact probability values are not required, the classification rule can be simplified:

$$y = \arg\max_{\hat{y} \in Y} P(x|y = \hat{y})P(\hat{y}) \tag{7}$$

Second, prior class probabilities, which might be utilized to incorporate user knowledge or a measure of data imbalance, are relatively easy to interpret. However, estimated class likelihoods mostly determine the classification decision. Each of the c class likelihoods are a product of d conditional feature probabilities, with c being the number of classes and d the number of dimensions. Unfortunately, these probabilities cannot be easily interpreted. This is the case because a high conditional feature probability does not necessarily indicate that a given feature is important for predicting a certain class and a low probability does not always mean that the feature is irrelevant. For example, let us consider feature x_0 of the motivating example. It is set in all instances, regardless of the class label, and thus has a high conditional feature probability for both the positive and the negative class. This feature provides no relevant information for classification, and the same applies to features that are never (or almost never) set in the training data, such as feature x_4. On the other hand, features x_1 and x_2 are always set in one of the classes and only once in the respective other class. In this case, the conditional feature probabilities of the active or inactive class, respectively, are three times higher than of the other one, which renders these two features highly descriptive.

A formal way to account for these feature probability relationships is provided by the so-called "odds ratio" (OR) of conditional probabilities:

$$\mathrm{OR}_d = \frac{P(x_d = 1|y = 1)}{P(x_d = 1|y = 0)} \tag{8}$$

The odds ratios for the motivating example are reported in Table 2. The approach of considering the odds ratios is theoretically established by rearranging the classification rule:

$$y = \arg\max_{\hat{y} \in Y} \prod_d P(x_d|y = \hat{y})P(\hat{y})$$
$$\Leftrightarrow y = \begin{cases} 1 & \text{if } \prod_d P(x_d|y = 1)P(y = 1) > \prod_d P(x_d|y = 0)P(y = 0) \\ 0 & \text{otherwise} \end{cases}$$
$$\Leftrightarrow y = \begin{cases} 1 & \text{if } \prod_d \frac{P(x_d|y = 1)}{P(x_d|y = 0)} > \frac{P(y = 0)}{P(y = 1)} \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

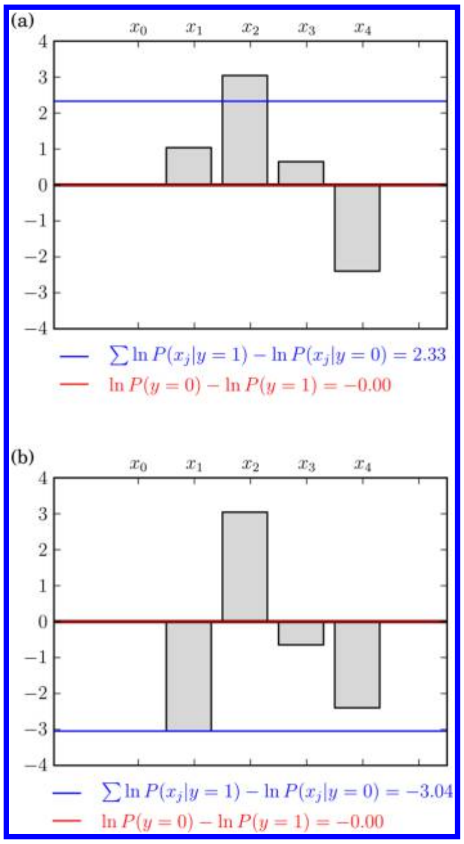


Figure 3. Principles of prediction visualization. The prediction for test instances (a) $z^{(0)}$ and (b) $z^{(1)}$ is visualized. Each bar represents the log odds ratio of a given feature for the test instance. Their sum is reported as a blue line and the difference between logarithmic prior probabilities as a red line.

The larger the odds ratio is, the higher is a feature's influence on the class likelihood; on the other hand, the smaller the odds ratios are, the smaller the class likelihood becomes. The closer the ratio is to 1, the less classification information is associated with the feature. Because two features with $\mathrm{OR}_i = x$ and $\mathrm{OR}_j = 1/x$ encode the same magnitude of classification information, a log transformation is applied such that $\log \mathrm{OR}_i = -\log \mathrm{OR}_j$ (cf. Table 2; in log space, features x_1 and x_2 have equal influence). The absolute value of the log odds ratio can then be regarded as a measure of an individual feature's importance for a given classification task. If it is negative, the presence of the feature is an indicator of the negative class label; if it is positive, it is an indicator of the positive class label.

Visualization. As rationalized in the previous section, a naïve Bayesian model contains only two parameters that must be learned for classification: the *class prior probabilities* (priors) and the *class likelihood*. For binary classification, there are two class priors that are related as follows:

$$P(y = 0) = 1 - P(y = 1) \tag{10}$$

However, there are four conditional probabilities for each dimension (feature) of the input space. For each pair of probabilities, eq 5 applies: $P(x_d = 0|y = \hat{y}) = 1 - P(x_d = 1|y = \hat{y})$.

If one would like to understand how a given model reaches a decision, one has to consider the odds ratio of each dimension in the input space.

Our primary goal is the visualization and interactive graphical analysis of a naïve Bayesian classification model with Bernoulli features. For *model visualization*, we introduce a scatter plot of its input dimensions using polar coordinates. The area of the plot is subdivided by features. Each point $p = (r,\theta)$ represents one dimension of the input space. Its radius is determined by the absolute value of its log odds ratio in the model, and the angles of all points are evenly distributed over the interval $[0,2\pi]$. Hence, the larger the distance between a point and the pole, the more important it is for classification. Coloring distinguishes features that indicate the positive class from features indicating the negative class, i.e., features with a negative log odds ratio are colored red and features with positive log odds ratio blue. Furthermore, points in the plot are scaled in size according to their maximum conditional probability for one of the classes:

$$s = \max\{P(x_d = 1|y = 0), P(x_d = 1|y = 1)\} \tag{11}$$

Therefore, it is possible to distinguish features occurring in most of the examples in one class from those occurring only in a few examples. The log odds ratio alone does not account for these different frequencies of occurrence.

In Figure 2, the model for our motivating example is visualized. It is evident that feature x_4 mostly determines the prediction, whereas features x_1−x_3 are less important and feature x_0, which maps to the pole, is not relevant. Furthermore, features x_2−x_4 (red) support negative class label predictions, whereas x_1 (blue) supports positive class label predictions (given its positive log odds ratio).

In addition to global model visualization, it is also possible to visualize and interpret individual predictions, which is relevant for assessing unexpected predictions and for model refinement.

For *prediction visualization*, one can exploit the fact that an instance obtains a positive class label if the following inequality applies (cf. eq 9):

$$\prod_d \frac{P(x_d|y = 1)}{P(x_d|y = 0)} > \frac{P(y = 0)}{P(y = 1)} \tag{12}$$

This inequality can also be expressed in log space:

$$\begin{aligned} &\log\left(\prod_d \frac{P(x_d|y = 1)}{P(x_d|y = 0)} > \frac{P(y = 0)}{P(y = 1)}\right) \\ &= \sum_d \log P(x_d|y = 1) - \log P(x_d|y = 0) \\ &> \log P(y = 0) - \log P(y = 1) \end{aligned} \tag{13}$$

Through the log transformation the product in eq 12 becomes a sum. Accordingly, a single prediction can be represented in a bar chart. In this chart, each bar represents a given feature, and the difference between the conditional probability given the active and inactive class is plotted. For *model visualization*, the conditional probability $P(x_d = 1|y = \hat{y})$ is utilized, as discussed above. However, for *prediction visualization*, we use the actual probability $P(x_d = z_d|y = \hat{y})$, with z being the example to be predicted. In addition, the sum of log probabilities is reported by a blue line and the sum of log priors by a red line. Hence, the final classification decision can be visualized in combination with the features that mostly influence the decision.

Figure 3 shows the prediction visualization for our motivating example. The priors for both classes are constant, but the class likelihood changes as a consequence of different input data. The test instances $z^{(0)}$ and $z^{(1)}$ differ in dimensions x_0, x_1, and x_3. Because x_0 has no impact on the classification (which can also be inferred from model visualization in Figure 2), it is not shown in Figure 3. For both instances, the fact that x_2 is not set in the training examples (Figure 1b) serves as an indicator for the positive class (i.e., it results in a positive odds ratio of x_2 in Figure 3), whereas the presence of x_4 is indicative of the negative class (i.e., it results in a negative odds ratio of x_4). Furthermore, features x_1 and x_3 in $z^{(0)}$ make small contributions to the overall class likelihood. Taken together, these probabilities result in a positive class label prediction for $z^{(0)}$. By contrast, in $z^{(1)}$, x_1 is not set and x_3 is set, which results in a large negative contribution to the likelihood for x_1 and a smaller negative contribution for x_3. As a consequence, the sum of log likelihoods falls below the sum of log priors. Accordingly, for $z^{(1)}$, the negative class label is predicted.

■ MATERIALS AND PROTOCOLS

Compound Data Sets. Three compound data sets of increasing complexity were used for Bayesian modeling and visualization. These data sets included two sets from ChEMBL (version 18),[33] i.e., carbonic anhydrase I inhibitors (CAI) and calcitonin gene-related peptide type 1 receptor ligands (CGRPR), and, in addition, a set of ATP-site directed inhibitors primarily focused on mitogen-activated protein kinase 14

Table 3. Compound Data Sets[a]

data set	no. of compds	no. of BMS	no. of CSK
CAI	1306	407	179
MAPK14 (active)	265	86	45
MAPK14 (inactive)	164	80	45
CGRPR	305	133	78

[a] For all three data sets, the number of compounds, unique Bemis-Murcko scaffolds (BMS), and corresponding carbon skeletons (CSK) is reported. For MAPK14, active and confirmed inactive compounds are listed separately. The other two data sets only consist of active compounds. In these cases, a random subset of ChEMBL was used as inactive compounds (see text). Scaffolds were calculated using OpenEye's OEChem toolkit.[40]

Table 4. Model Performance[a]

	no. of training compds		no. of test compds				
data set	active	inactive	active	inactive	precision	recall	F1-score
CAI	1044	8000	262	2000	0.5360	0.9084	0.6742
MAPK14	212	131	53	33	0.8200	0.7736	0.7961
CGRPR	244	8000	61	2000	0.5495	1.0	0.7093

[a] For each set, the number of active and inactive training instances used to build a naïve Bayesian classification model, and the number of active and inactive test compounds are given. In addition, the classification performance is reported. Precision is calculated as the ratio of true active predictions over all active predictions, recall is the ratio of correctly predicted actives over all actives, and the F1-score is the harmonic mean of both.

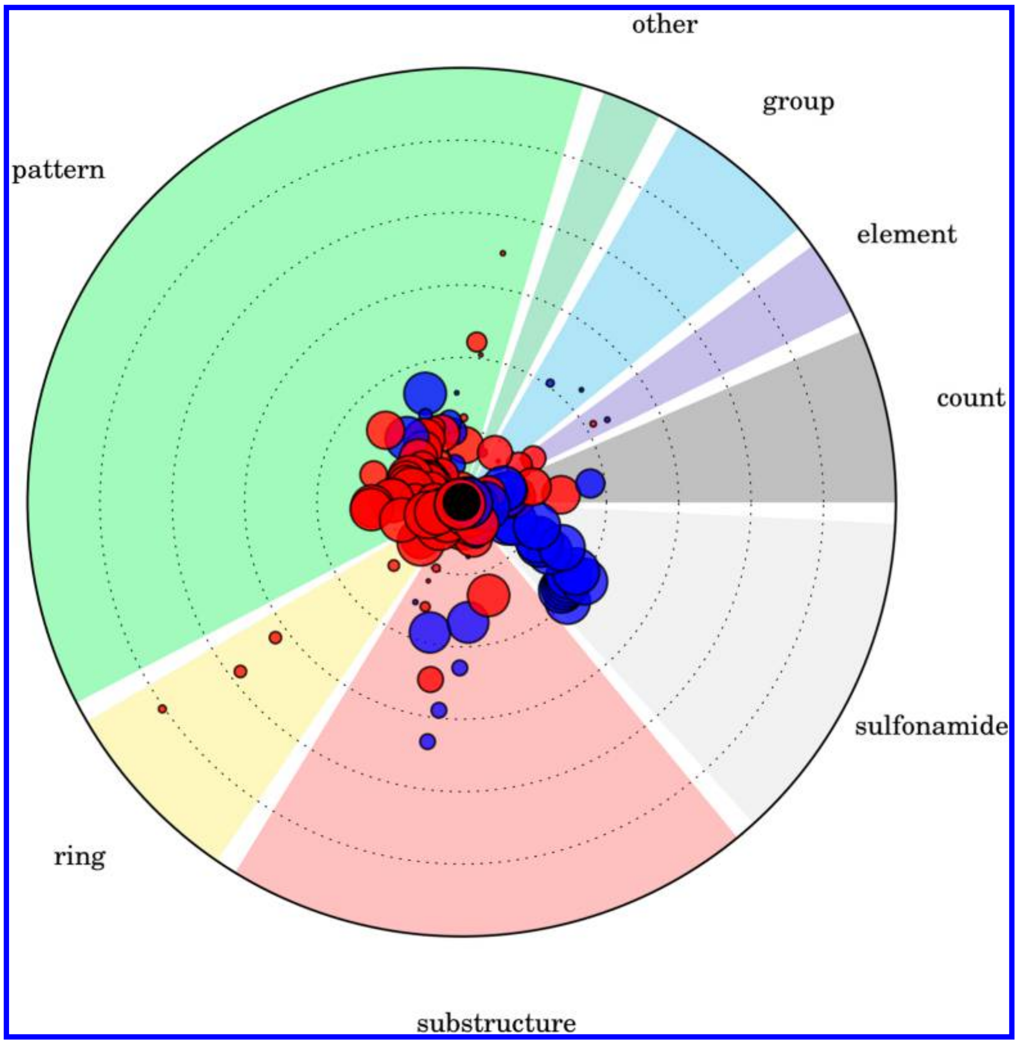


Figure 4. CAI model visualization. All 22 MACCS features associated with the signature sulfonamide are combined into one group. Graphical analysis confirms the hypothesis that the classification model is primarily emphasizing sulfonamide-associated features (see text for further details).

(MAPK14)[34] originating from the ProQinase free choice biochemical assay system.[35] The composition of the data sets is reported in Table 3.

Selected ChEMBL compounds were required to be tested in a direct binding assay with a ChEMBL confidence score of 9. Furthermore, only compounds were considered for which a K_i value of less than 10 μM was available. Using in-house scripts, all candidate compounds were filtered for duplicates, undesired reactive groups, and PAINS[36] liabilities to arrive at a final selection. From compounds in all three data sets, Bemis-Murcko scaffolds (BMS)[37] and corresponding carbon skeletons (CSK),[38] in which all heteroatoms are converted to carbons and all bond orders set to 1, were systematically extracted. Decreasing compound-to-BMS and compound-to-CSK ratios generally indicate increasing structural diversity.

The CAI set was selected because it contained small inhibitors mostly sharing a sulfonamide group (992 of 1306 compounds), which is a hallmark for carbonic anhydrase inhibition. Hence, in this case, a defined chemical moiety was known to be a major determinant of activity. This inhibitor set yielded 407 BMS, 179 CSK, and compound-to-BMS and -CSK ratios of 3.21 and 7.30, respectively.

The MAPK14 set consisted of 429 ATP-site directed kinase inhibitors, 265 of which inhibited the MAP14 kinase. Hence, the remaining compounds were confirmed to be inactive against this kinase (but, in part, active against other kinases). All 429 inhibitors contained a conserved pyridinyl-imidazole core with varying substitutions. Hence, this data set was selected as a structurally homogeneous set enabling the derivation of predictive models of MAP14 kinase activity.

The CGRPR set was characterized by a much higher degree of structural diversity than the other two sets, with compound-to-BMS and -CSK ratios of 2.29 and 3.91, respectively, and was therefore selected for our analysis. It should be noted that this compound set had the lowest ratios among 122 structurally heterogeneous candidate sets extracted from ChEMBL for which we have internally built and evaluated naïve Bayesian classification models.

For classification of CAI and CGRPR ligands, a random sample of 10,000 other compounds was taken from ChEMBL to serve as negative training and test instances. For the kinase data set, confirmed inactive compounds were used.

Molecular Representation. MACCS structural keys[39] were used as an exemplary molecular representation. The public version of the MACCS fingerprint consists of a set of 166 structural fragments or patterns, which were generated using an in-house program based upon OpenEye's OEChem toolkit[40] and SMARTS patterns adapted from RDKit.[41] For visualization, MACCS features were organized into different groups:

1. "ring": All ring-related features, e.g., "4M ring" (13 MACCS features),
2. "count": Occurrence count features, e.g., "O > 2" (13 features),
3. "group": Features representing a periodic table group, e.g., "actinide" (11 features),
4. "element": Features representing single specific element, e.g., "P" (9 features),
5. "substructure": Specific substructures, e.g., "ON(C)C" (39 features),
6. "pattern": Substructures with wildcards or exclusions, e.g., "QAAA@1" (76 features),
7. "other": All remaining features (5 features).

Features potentially falling into multiple categories were assigned to a single group in the order of decreasing priority from groups 1−6. For example, the feature "Aromatic Ring >1" was assigned to the "ring" group.

The visualization is also applicable to other types of binary fingerprint representations such as fragment or extended connectivity fingerprints. In the current study, we limit the application to the MACCS fingerprint because of its small size and ease of interpretation. A prototypic implementation of the visualization method is made available (see below), which also provides a basis for further studies with other molecular representations.

Model Building and Evaluation. Models were generated as described in the Concepts and Methods section using the naïve Bayesian formulation with Bernoulli features of the freely available Python machine learning toolkit Scikit-learn.[42] A smoothing factor $\alpha = 1$ and example weights inversely proportional to the class balance were used, which prevented potential smoothing artifacts due to imbalanced data and resulted in assumed uniform prior probabilities. As reported in Table 4, each model was trained on a random subset of 80% of the active compounds. The remaining 20% were used as positive test instances. For CAI and CGRPR, 8000 and 2000 randomly chosen ChEMBL compounds were used as negative training and test examples, respectively. For MAPK14, 80% and 20% of the confirmed inactive compounds were used as negative training and test examples, respectively (Table 4). Hence, the composition of training and test sets for MAPK14 modeling principally differed from CAI and CGRPR.

RESULTS AND DISCUSSION

The principles of *model visualization* and *prediction visualization* are illustrated in Figure 2 and Figure 3, respectively, and have been discussed in the Concepts and Methods section. In the following, we present a number of data set applications to evaluate the visualization techniques in greater detail and analyze models and predictions. First, the prediction performance of the different Bayesian classification models is reported.

Model Performance. In Table 4, the performance of the naïve Bayesian classifiers derived for the three compound data sets is summarized. For CAI and CGRPR from ChEMBL, recall is very high (~0.91 and 1.00, respectively) but precision only intermediate (~0.54 and ~0.55, respectively). For MAPK14, a smaller and more balanced data set, recall performance is lower (~0.77) but precision higher (0.82) than for the ChEMBL data models, which results in a higher F1-score (~0.80). Overall, the classifiers derived for compound data sets of different composition and structural complexity display reasonable to high accuracy, a prerequisite for meaningful evaluation of classification models and predictions.

Model Visualization. *CAI.* In Figure 4, the naïve Bayesian model for CAI is visualized. In this case, an additional feature group was defined to which the 22 MACCS features were assigned that are associated with the sulfonamide substructure "*S(=O)(=O)N". Thus, in the scatter plot, all features related to the sulfonamide group are easily identified. Blue coloring of these features confirms that the classification model associates features set in sulfonamide-containing inhibitors with activity. In addition, the size of the corresponding feature points indicates that these features are set in most of the active compounds. While 115 of the 166 MACCS features have an absolute log odds ratio smaller than one, 13 of the 22 sulfonamide features have an

Table 5. Log Odd Ratios and Class Likelihoods of Selected CAI Model Features[a]

feature	group	log odds ratio	class likelihood
4 M ring	ring	−5.0271	0.0169
3 M ring	ring	−3.8505	0.0502
QAAA@1	pattern	−3.4860	0.0036
OS(O)O	substructure	3.3418	0.0844
7 M ring	ring	−3.1828	0.0488
S−O	substructure	2.8918	0.0853
NC(C)N	substructure	−2.4876	0.2433
Si	element	2.3116	0.0049
OQ(O)O	substructure	2.2906	0.0863
group IVa,Va,VIa rows 4−6	group	2.2683	0.0011
C=C(Q)Q	pattern	−2.2216	0.1423
P	element	−2.1154	0.0089
QAA@1	pattern	−2.0519	0.0009
group IIIA (B...)	group	2.0500	0.0164
QQH	sulfonamide	2.0193	0.8227
NS	sulfonamide	2.0088	0.7787

[a] Reported are all MACCS features from the CAI prediction model having an absolute log odds ratio greater than two. If the log odds ratio is negative, the class likelihoods are reported for the negative class; if the log odds ratio is positive, they are reported for the positive class. Log odds ratios and class likelihoods reflect the radius and size of the feature points in Figure 4.

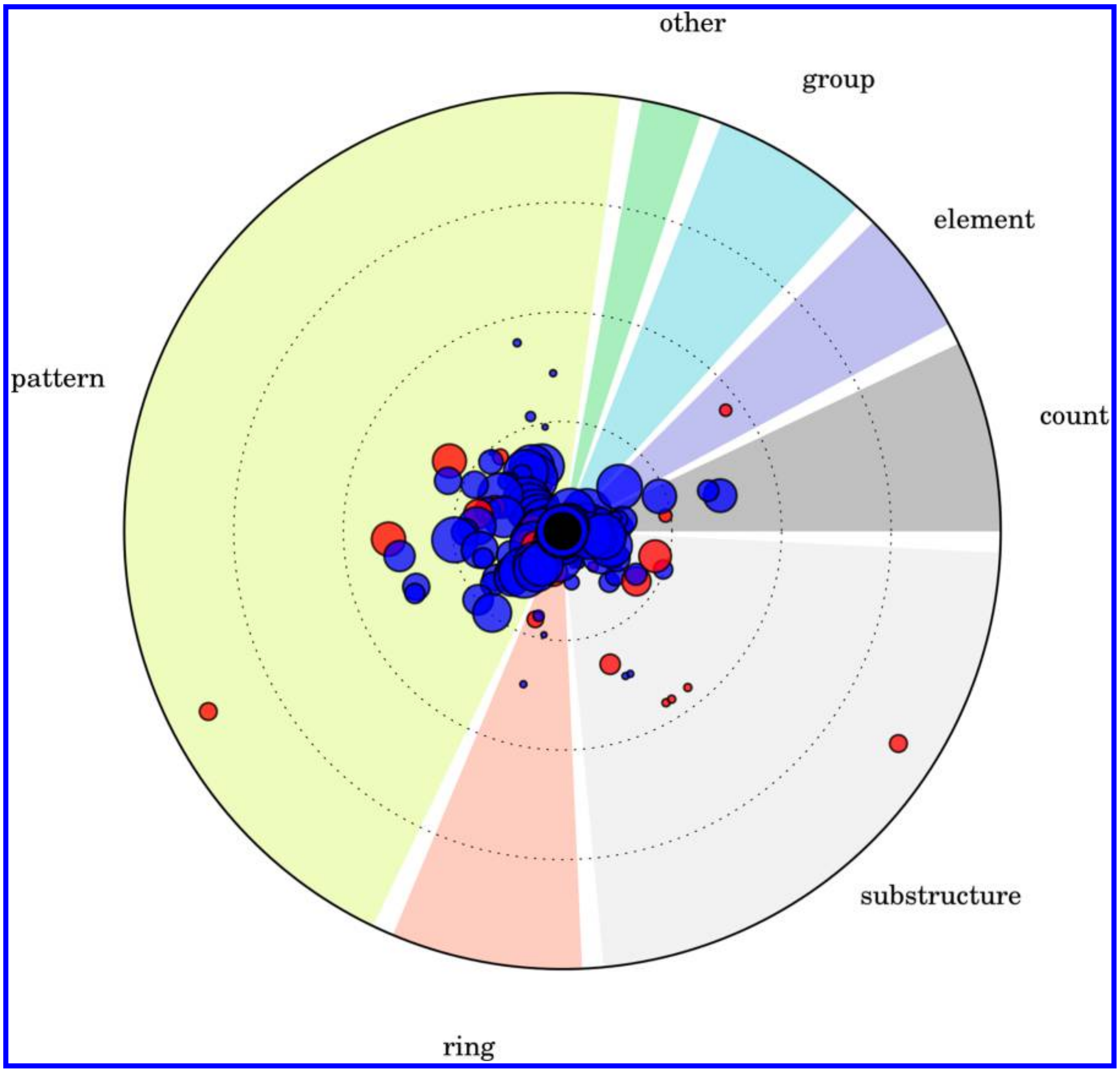


Figure 5. MAPK14 model visualization. Features important for selection of active and deselection of inactive compounds are identified.

absolute log odds ratio larger than one, which confirms their relevance for the model. However, the visualization also reveals that there are other features with much larger absolute log odds ratios than the sulfonamide features, which include three "ring" and three "pattern" features indicating inactivity, two features from the "element" group (one promoting activity and the other inactivity), and four features from the "substructure" group (three indicating activity and one inactivity). All features having an absolute log odds ratio greater than two are summarized in Table 5. However, while the log odds ratio of these features is high, their class likelihood is rather low, as reflected by smaller points. This is indicative of an underrepresentation of these features in the data. Exceptions include the substructure NC(C)N and the pattern C═C(Q)Q, which are present in 24.3% and 14.2% of the inactive compounds, respectively, and have log odds ratios of −2.49 and −2.22. This means that these features are approximately 12 and 10 times more likely to appear in inactive than active compounds. Furthermore, the two sulfonamide features QQH and NS are contained in 82.3% and 77.9% of the active compounds, respectively, and have a log odds ratio of 2.02 and 2.01, indicating that they are approximately 7.5 times more likely to appear in active than inactive compounds.

Taken together, the visualization of the CAI model clearly not only confirms a critically important role of the sulfonamide moiety for the prediction of activity but also demonstrates that there are other features the model regards as even more important for prediction of activity than the sulfonamide. The relatively small class likelihoods of these features point at data imbalance during training, consistent with the limited precision of the model.

MAPK14. The visualization of the MAPK14 model is shown in Figure 5. Here, we again observe that most of the MACCS features have an absolute log odds ratio smaller than one and can hence be considered less important for activity prediction. However, there are a number of features with absolute log odds ratios between one and two, and most of these features promote activity. Finally, a substructure and a pattern feature (C═N and N=A) have absolute log odds ratios larger than three and thus elicit the largest influence on the prediction of activity, provided they are set in the fingerprint of a given compound. Interestingly, both features have a negative log odds ratio, meaning that they are used by the model to deselect inactive compounds, rather than select active ones. By contrast, the features "QCH2A > 1" and "CH3AACH2A" support prediction of activity. All features with an absolute log odds ratio greater than one are listed in Table 6.

Table 6. Log Odd Ratios and Class Likelihoods of Selected MAPK14 Model Features[a]

feature	group	log odds ratio	class likelihood
C=N	substructure	−3.6285	0.1091
N=A	pattern	−3.6285	0.1091
BR	element	−1.8536	0.0484
S−O	substructure	−1.8304	0.0181
OS(O)O	substructure	−1.8304	0.0181
OQ(O)O	substructure	−1.8304	0.0181
CH2=A	pattern	1.7671	0.0170
A$A($A)$A	pattern	−1.5894	0.4279
CH3AAACH2A	pattern	1.5042	0.3546
QCH2A > 1 (&...)	count	1.4754	0.4109
CH3AACH2A	pattern	1.4627	0.1436
S heterocycle	ring	1.4436	0.0123
NS	substructure	1.4436	0.0123
CSN	substructure	1.4436	0.0123
C=C(Q)Q	pattern	1.4436	0.0123
CH3ACH2A	pattern	1.4289	0.2655
CH3 > 2 (&...)	count	1.3799	0.1623
N−O	substructure	−1.2919	0.1471
NAAAN	pattern	−1.2141	0.4203
CH2QCH2	pattern	1.1404	0.2702
QHAQH	pattern	1.0849	0.0310

[a] Reported are all MACCS features from the MAPK14 prediction model having an absolute log odds ratio greater than one. Log odds ratios and class likelihoods reflect the radius and size of the feature points in Figure 5.

CGRPR. The CGRPR model is visualized in Figure 6. For this structurally diverse compound set, the visualization of the model notably differs from the others. Here, all of the features that promote activity have a log odds ratio smaller than two, whereas the absolute value of the negative log odds ratios even exceeds six, which corresponds to a more than 400 times higher class likelihood for inactive over active compounds. The features with the highest positive log odds ratios are NC(O)N, the 7-membered ring, and S−S with ORs of 1.85, 1.78, and 1.58, respectively. This corresponds to a class likelihood for active compounds that is 4.8−6.3 times higher than for inactive ones. However, the probability of the count feature "QQ > 1" to be set in the negative class is 429 times higher than its probability to be set in the positive class, and similar values are observed for substructures N−O and NO. Features with an absolute log odds ratio of more than three are summarized in Table 7. It can be seen that features with absolute log odd ratios greater than four occur in only a small fraction of randomly chosen ChEMBL compounds. However, there are also features such as OAOO of A$A!S that appear in more than 10% of all assumed negative instances.

The Bayesian classification model of CGRPR successfully not only recovers all active test compounds but also has only intermediate precision. Visualization of the model reveals that nearly all features with a large absolute log odds ratio promote the prediction of inactivity, provided they are present in a compound. This indicates that the model primarily deprioritizes inactive compounds instead of prioritizing active ones. This observation is consistent with the fact that many active compounds in this data set are structurally diverse and cannot be easily distinguished from negative instances by only a few descriptive features. Instead, the model focuses on features that predominantly occur in inactive compounds.

Table 7. Log Odd Ratios and Class Likelihoods of Selected CGRPR Model Features[a]

feature	group	log odds ratio	class likelihood
QQ > 1 (&...)	count	−6.0607	0.0520
N−O	substructure	−6.0139	0.0496
NO	substructure	−6.0139	0.0496
4 M ring	ring	−4.9351	0.0169
QHQH (&...)	pattern	−4.9277	0.0167
QCH2Q	pattern	−4.8661	0.0157
P	element	−4.2923	0.0089
OQ(O)O	substructure	−4.2781	0.0087
I	element	−4.2637	0.0086
CH2=A	pattern	−3.9807	0.0065
OAAO	pattern	−3.7200	0.1741
ON(C)C	substructure	−3.7182	0.0050
A$A!S	pattern	−3.5099	0.1411
CQ(C)(C)A	pattern	−3.4631	0.0039
QAAA@1	pattern	−3.3963	0.0036
NS	substructure	−3.2094	0.1045
OS(O)O	substructure	−3.2069	0.0030
CSN	substructure	−3.1716	0.1006
QHAAAQH	pattern	−3.1219	0.0957

[a] Reported are all MACCS features from the CGRPR prediction model having an absolute log odds ratio greater than three. All log odds ratios are negative indicating that all features support prediction of inactivity.

Model Characteristics. Taken together, model visualizations for the three compound data sets reveal the presence of different model characteristics. The CAI model primarily prioritizes compounds containing the sulfonamide signature (as to be expected) and deprioritizes compounds with specific ring systems or patterns. The MAPK14 selects compounds on the basis of specific features that are preferentially set in active compounds. By contrast, the CGRPR model primarily deselects inactive compounds instead of prioritizing actives. Hence, classification models derived for data sets of varying composition and structural complexity display different model characteristics. Graphical analysis clearly reveals key feature for predictions using the different models.

Feature Selection. Another interesting application of model visualization is the rationalization of feature selection effects. To illustrate this point, we have subsequently removed the features with the highest log ORs from our models and monitored the change in model performance (data not shown). Removal of the first few features from the CAI model did not alter performance significantly. This might seem surprising at first glance. However, Figure 4 shows that features with highest log OR were only very infrequently set. Therefore, removal of these features did not influence the majority of new predictions. By contrast, when features with a higher probability to be set were removed, for instance, the larger circles from the "substructure" and "sulfonamide" areas, classification performance changed significantly. Equivalent observations were made for the MAPK14 model. Removal of features with high log odd ratios only slightly affected predictive performance, if these features were only rarely set. Removal of additional features resulted in further improved recall but reduced precision—a direct consequence of predicting more compounds as active. This effect can also be rationalized by analyzing Figure 5 where the outermost features were responsible for compound deselection. Finally, removing the most important features from the CGRPR model resulted in reduced precision, which can also be attributed to the fact that

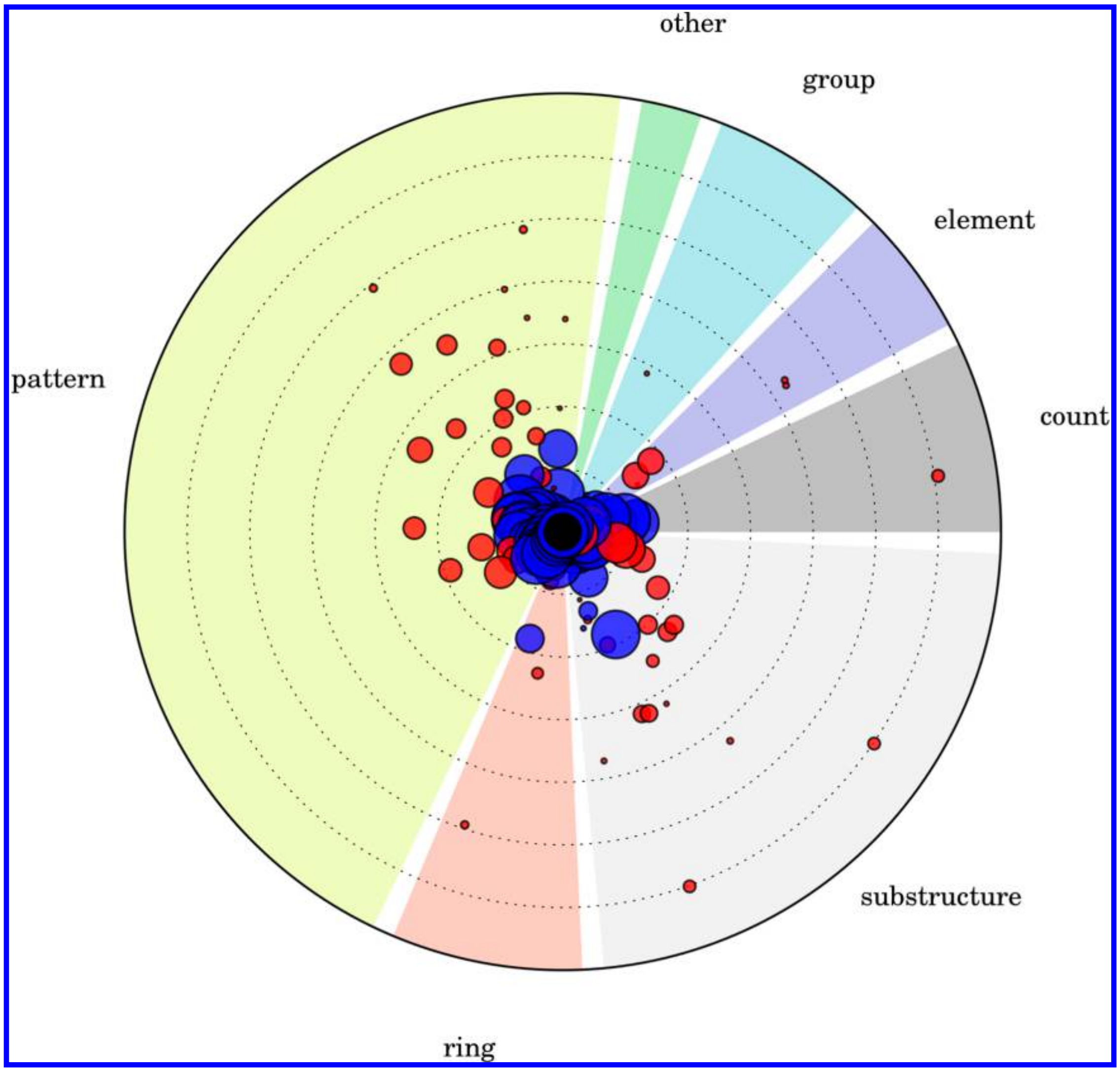


Figure 6. CGRPR model visualization. The prediction model for the structurally diverse CGRPR class strongly deselects inactive compounds based on features with negative log odd ratios (red circles).

only deselecting features were removed. Clearly, overall prediction performance is predominantly affected by distinguishing features that are frequently set in compounds, which can be well appreciated on the basis of model visualization.

Prediction Visualization. It is important to note that the model visualization emphasizes log odds ratios given features are set in compound fingerprints, i.e., $P(x_d = 1|y = 1)/(P(x_d = 1|y = 0))$ (cf. eq 8). However, if a feature is not set in the fingerprint of a compound, the term entering the classification rule is changed to

$$\frac{P(x_d = 0|y = 1)}{P(x_d = 0|y = 0)} = \frac{1 - P(x_d = 1|y = 1)}{1 - P(x_d = 1|y = 0)} \qquad (14)$$

This means that features with a high log odds ratio given their presence can have low log odds ratios in their absence, which can further complicate the understanding of model decisions. Hence, to better understand individual predictions, rather than global model performance, a *prediction visualization* method has also been introduced.

Table 8 summarizes true and false positive and negative predictions for the test sets of the three models. We will use compounds from these different subsets as examples for prediction visualization.

Table 8. True or False Positive and Negative Predictions[a]

data set		no. of cpds predicted to be active	no. of cpds predicted to be inactive
CAI	no. of active compounds	238	24
	no. of inactive compounds	206	1794
MAPK14	no. of active compounds	41	12
	no. of inactive compounds	9	24
CGRPR	no. of active compounds	61	0
	no. of inactive compounds	50	1950

[a] For each model, the number of true positives (correctly predicted active compounds), true negatives (correctly predicted inactive compounds), false positives (inactive compounds predicted to be active), and false negatives (active compounds predicted to be inactive) is reported.

CAI. Exemplary true and false positive predictions of the CAI model are visualized in Figure 7. In the fingerprint of the correctly predicted active compound, 53 of 166 features are set, 33 of which have a positive and 20 a negative log odds ratio. However, 42 and 71 of the 113 MACCS substructures not present in the fingerprint have a positive and negative log odds ratio, respectively. Considering eq 14, this means that the 71 features that are not set and have a negative log odds ratio in the

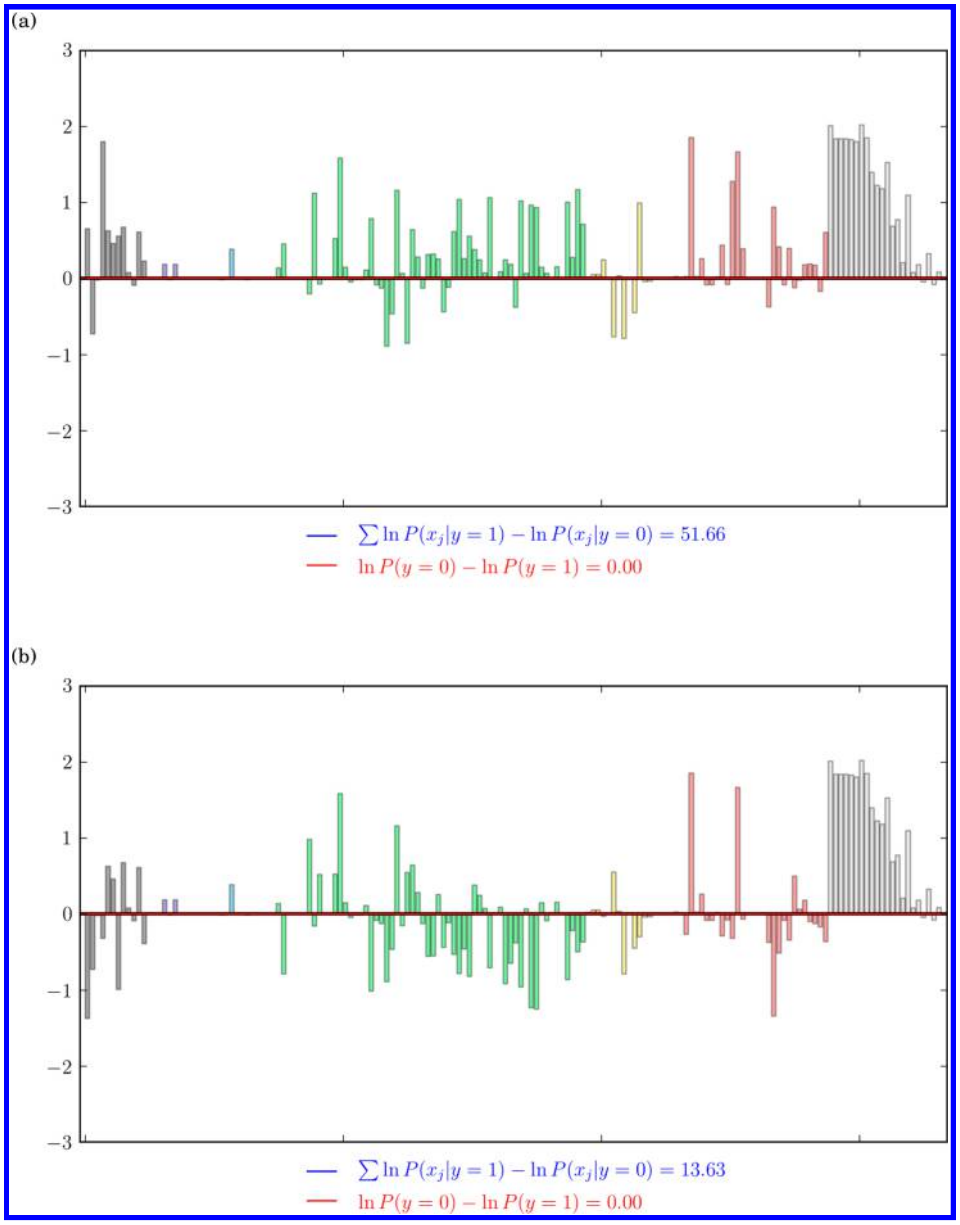


Figure 7. Prediction visualization for CAI. Shown is the prediction visualization for two compounds predicted to be active and representing a (a) true positive and (b) false positive, respectively.

global model actually make a positive contribution to the prediction of activity for this compound. In total, the presence or absence of 104 (33 plus 71) of 166 MACCS features contributes to the prediction of activity, whereas the remaining 62 support prediction of inactivity. The actual weight of their respective contributions is depicted in Figure 7a. It is evident that most of the negative contributions are small (none of them exceeds −1). On the other hand, many of the positive contributions fall in the range between +1 to +2. In total, the sum of positive log class likelihoods is 51.66, which results in a clear positive prediction, considering that the smallest ratio of class log priors to be exceeded for a positive prediction is 0. Taking the group coloring in Figure 4 into account, one can immediately conclude that the most positive contributions result from the sulfonamide group (light gray), followed by the pattern (green), the substructure (red), and the count group (dark gray). By contrast, significant negative contributions come from the feature "Heterocyclic atom >1" (count group; dark gray), the ring features "5 M ring" and "N Heterocycle" (yellow), and the patterns "NAAN" and "QAAAA@1" (green). According to the model, these features are more likely to be set by inactive compounds, but the presence or absence of features that support prediction of activity outweighs these contributions.

Figure 7b shows the visualization for a false positive prediction by the CAI model. In this case, there are more negative contributions from different feature groups than for the example in Figure 7a. Yet, the sum of log odd ratios still is 13.63, giving rise to a positive prediction. The features with the largest influence on this prediction include the patterns "A$A!S" and "SA(A)A" (green), the substructures "CSN" and "CSO" (red), and most of the features associated with the sulfonamide group (light gray). The latter contributions are largely responsible for the false positive prediction, although other infrequently observed structural features render this compound inactive. This reflects a limitation of the model for predicting compounds that contain the sulfonamide but are nonetheless inactive for other reasons

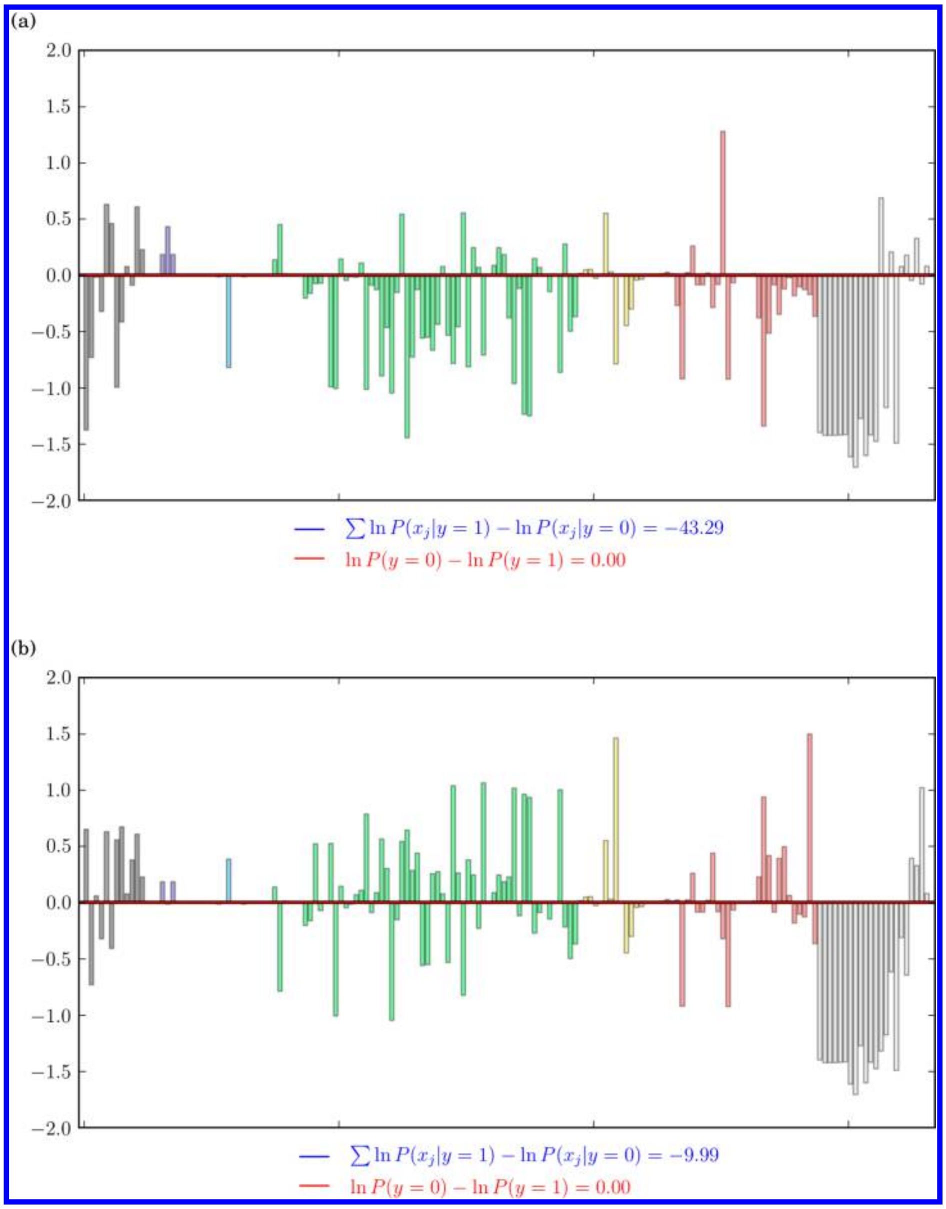


Figure 8. Prediction of inactivity for CAI. Shown is the prediction visualization for two compounds predicted to be inactive and representing a (a) true negative and (b) false negative, respectively.

(such as, for example, steric hindrance due to the presence of other groups).

Exemplary true and false negative CAI predictions are visualized in Figure 8. For the correctly predicted inactive compound, many negative contributions are observed resulting in a sum of class log likelihoods of −43.29 (Figure 8a). By contrast, the active compound has a more balanced ratio of both positive and negative contributions resulting in a sum of class log likelihoods of −9.99 (Figure 8b). Although many positive contributions are detected for the active compound, the negative prediction is mostly due to the absence of a sulfonamide group in this compound, which again reflects the focus of the classifier on this signature group shared by the majority of active compounds. Features associated with this group (light gray) make the strongest negative contribution due to their absence.

MAPK14. Figure 9a visualizes a true positive prediction by the MAPK14 model. Here, both positive and negative contributions of the count group (dark gray), the pattern group (yellow), and the substructure group (light gray) become apparent. Features resulting from single elements (purple), groups (blue), or rings (red) only make minor positive contributions to the class likelihood. In total, however, the positive contributions outweigh the negative ones; hence, the compound is predicted to be active. The largest positive contributions come from the presence of the patterns “QHAQH”, “S=A”, and “QA(Q)Q” and the largest negative terms from the absence of the patterns “AQ(A)A” and “QCH2A”. Overall, this prediction clearly reflects the presence of a cumulative effect of many small-magnitude contributions accounted for by the model. Figure 9b visualizes a true negative prediction using this model, which helps to better understand its strong tendency to deselect inactive compounds, as discussed above. There are positive contributions from count features, substructures, and patterns, but the magnitude of negative contributions is by far larger. The absence of a sulfur atom (purple) and of the patterns “A!N$A”, “QA(Q)Q”, and “AN(A)A” (yellow) have the strongest negative influence. In fact, most influential terms come from the absence of substructural features. Hence, the finding from the global

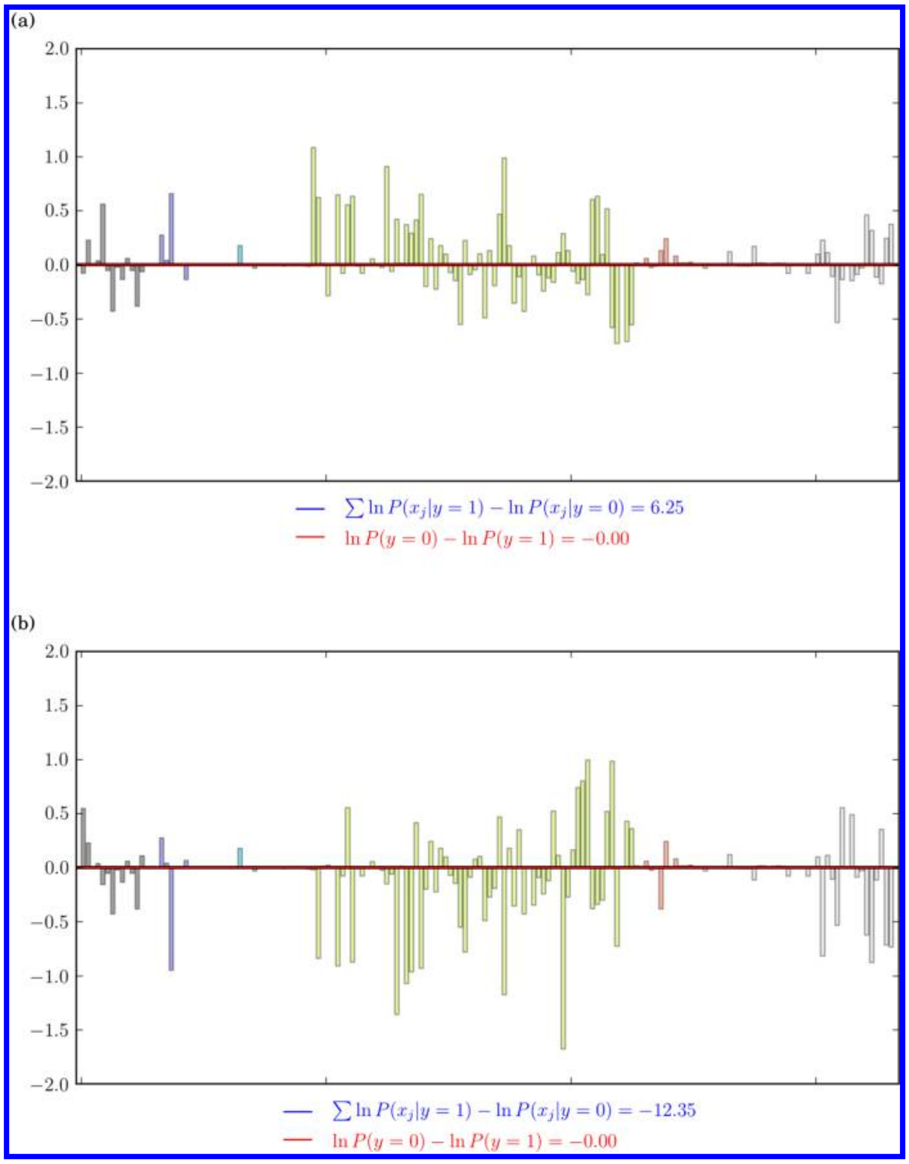


Figure 9. Correct predictions for MAPK14. Shown is the prediction visualization for two compounds representing a (a) true positive and (b) true negative, respectively.

model analysis that the MAPK14 model preferentially deselects inactive compounds is further substantiated at the level of individual predictions.

CGRPR. Figure 10 analyzes active and inactive compounds correctly predicted by the CGRPR model, respectively. For the true positive prediction in Figure 10a, there are a four major negative terms in the log likelihood sum accounting for the absence of the pattern "C$=C($A)$A" and the substructure "C═CN" and for the presence of the patterns "QHAAQH" and "QHAQH". On the other hand, there are two peaks indicating positive contributions including the 7-membered ring (red) and the substructure "NC(O)N" (light gray), which are both set in the fingerprint of this compound. The remaining positive contributions are comparably small. Hence, in this case, there also is a cumulative effect leading to the prediction of activity for this compound. This is consistent with the structural heterogeneity of the CGRPR set and the absence of simple structural rules that distinguish active from inactive compounds (such as the presence or absence of specific functional groups).

In Figure 10b, a completely different picture emerges. Here, the majority of features in the fingerprint have a positive log odds ratio, although all of these contributions are of rather low magnitude. In addition, there are a few small and medium-size contributions to negative class log likelihood, with three major feature peaks. These include the absence of patterns "NAN" and "QHAAACH2A" and of the substructure "NH". In other words, the model does not strictly require any active compound to possess these features but strongly deprioritizes compounds that do not have these features set in their fingerprints.

The visualization of the inactive prediction in Figure 10b also illustrates another important aspect: While the model visualization in Figures 4−6 can highlight features that (according to the model) make significant contributions to the prediction of activity or inactivity, it cannot fully represent the decision process for any new test instance. For example, Figure 6 indicates that features "QQ > 1″, "N−O", and "NO" might dominate the predictions. However, this is only the case if these features are present in a test compound. The prediction for compounds where these features are not set, as the one depicted in

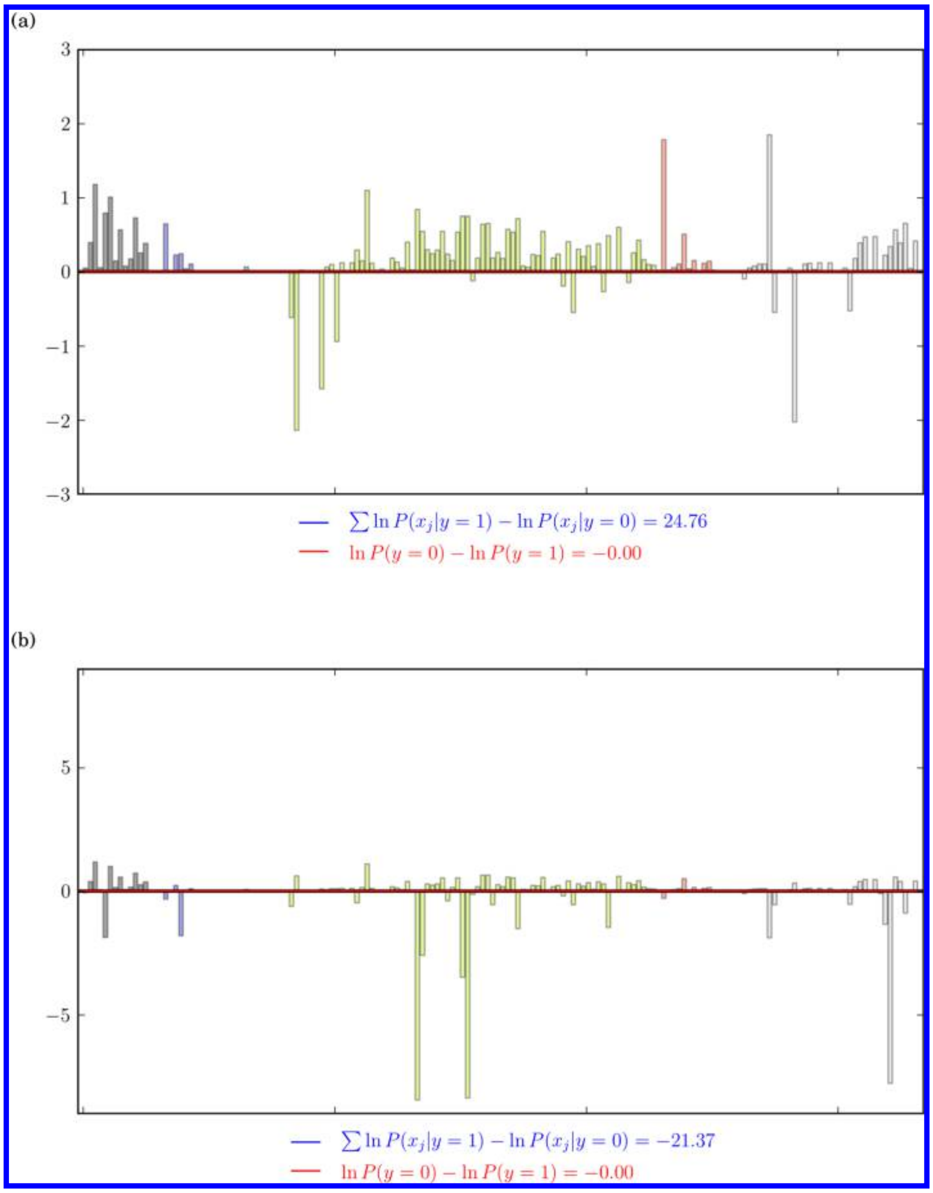


Figure 10. Correct predictions for CGRPR. Shown is the prediction visualization for two compounds representing a (a) true positive and (b) true negative, respectively.

Figure 10b, is hardly influenced by any of them. For example, let us consider the feature "QQ > 1". If it is present, its log odds ratio is given by

$$\log \frac{P(x_{\mathrm{QQ}>1} = 1|y = 1)}{P(x_{\mathrm{QQ}>1} = 1|y = 0)} = \log \frac{0.0001}{0.0520} = -6.0607$$

However, if it is absent, the log odds ratio approaches zero:

$$\log \frac{P(x_{\mathrm{QQ}>1} = 0|y = 1)}{P(x_{\mathrm{QQ}>1} = 0|y = 0)} = \log \frac{1 - P(x_{\mathrm{QQ}>1} = 1|y = 1)}{1 - P(x_{\mathrm{QQ}>1} = 1|y = 0)}$$

$$= \log \frac{0.9999}{0.9480} = 0.0533$$

In this case, other features play a by far more important role. It follows that both a careful analysis of global model performance as well as of individual predictions is required to fully rationalize why classification models might yield—or not yield—accurate predictions.

Feature Mapping. Depending on the chosen molecular descriptors, features that are most important for a prediction can be back-projected onto test compounds, which aids in the exploration of SARs. Fragment fingerprints such as MACCS are suitable descriptors for feature mapping. Figure 11 shows examples of correctly predicted active and inactive compounds and of key features that are present in these compounds and make major contributions to the prediction of activity or inactivity. For each compound, only features with a log odds ratio of at least 90% of the maximum or minimum OR are mapped. These features are reported in Table 9. For example, features of the compound in Figure 11a have a minimum and maximum log odds ratio of −0.89 and 2.02, respectively. Requiring at least 90% of this value, we highlight features with log odds ratios smaller than 0.8 or greater than 1.82. By contrast, no color-coding can be applied for the compound in Figure 11b because it is predicted to be inactive since it is missing essential substructures. This illustrates limitations of feature mapping approaches. The examples in Figure 11 reveal that features that are present in

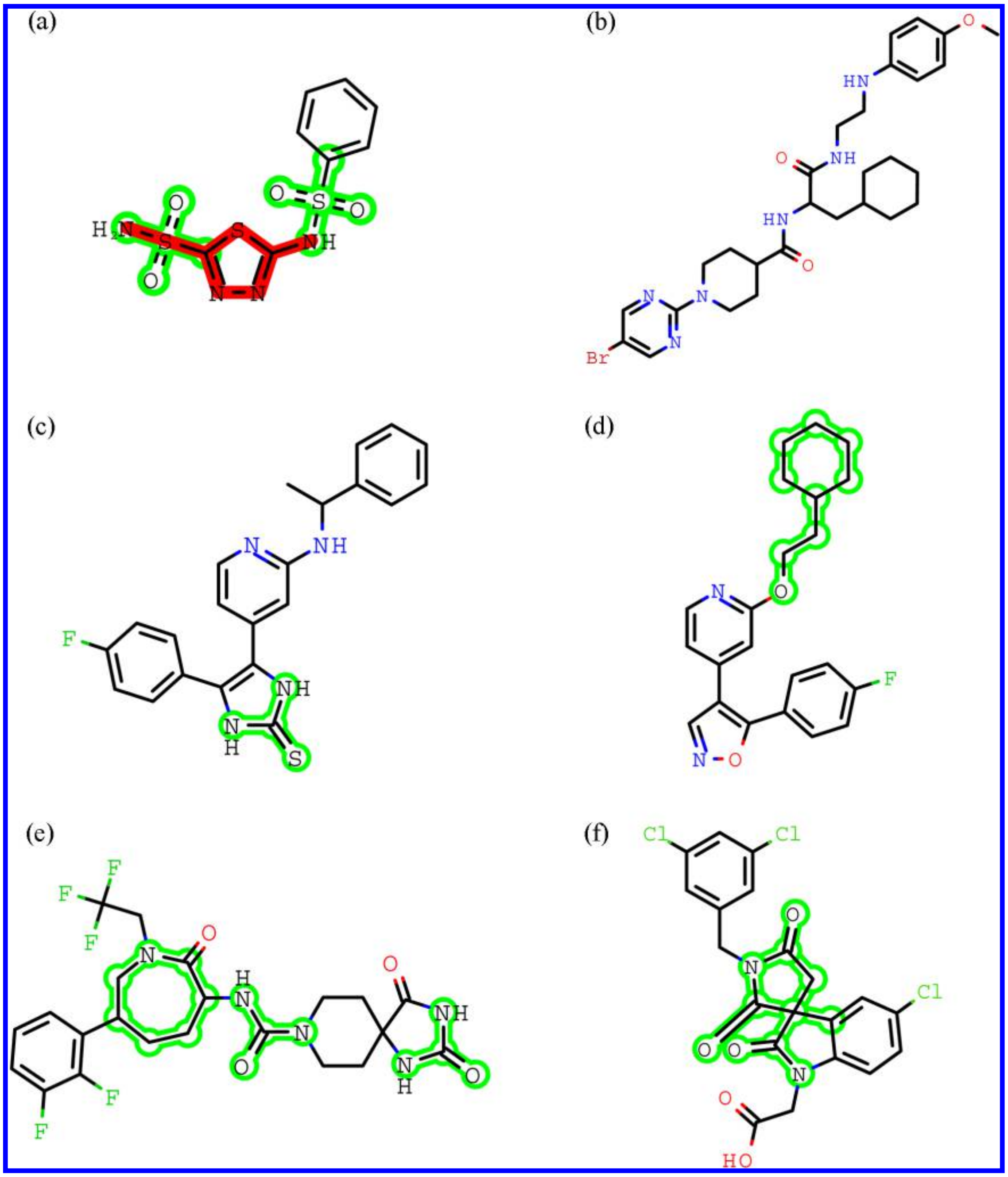


Figure 11. Feature mapping. Features with at least 90% log odds ratio compared to the maximum or minimum are back-projected onto test compounds. Shown are correctly predicted active and inactive compounds for which predictions are visualized in Figures 7−10: (a) active/CAI, (b) inactive/CAI, (c) active/MAPK14, (d) inactive/MAPK14, (e) active/CGRPR, and (f) inactive/CGRPR. Red and green coloring indicates features with negative and positive log odd ratios, respectively. The depiction of the feature mapping was created using OpenEye's OEDepict Toolkit.[43]

compounds and determine predictions cover substructures of different size. However, feature mapping only provides an incomplete account of predictions, in contrast to prediction visualization, because of the frequent importance of feature absence, as revealed in our analysis.

Figure 12 shows two active and two inactive compounds from the CAI test set. The two compounds at the top were correctly predicted as active, whereas the bottom left compound represented a true negative and the bottom right compound a false positive prediction. Color shading indicates the magnitude of the mapped features' log odds ratio (only features with a log OR of at least 25% of the maximum absolute log OR were considered for mapping). Interestingly, the compounds predicted to be active also contain a rather large red area, i.e., features that the model utilized to deselect inactives. However, nonset features that cannot be mapped often had a major influence on the predictions. For instance, the sum of log ORs of the features that are set in the compound in Figure 12a is −0.14. This means that this compound would be predicted to be inactive if the prediction would only be based on mapped substructures. However, the sum of log ORs of the nonset features is 2.52, which hence leads to the prediction of activity. The other two compounds predicted to be active have a log OR sum of 5.63 and 23.94 for their set features, and 6.77 and 22.08 for their nonset features, respectively. The true negative prediction in Figure 12c has an overall log OR sum of −3.8 (with both the sum of the set

Table 9. Features Most Important for Individual Predictions[a]

compound	feature	presence	OR
(a)	CSN	+	1.85
	NS	+	2.01
	OSO	+	1.84
	QSQ	+	1.84
	S=O	+	1.84
	AS(A)A	+	1.82
	QQH	+	2.02
	S=A	+	1.85
	NAAN	+	−0.89
	QAAAA@1	+	−0.85
(b)	C=C(C)C	-	1.28
	QQH	-	−1.61
	S=A	-	−1.71
	S	-	−1.60
(c)	QHAQH	+	1.08
	QA(Q)Q	+	0.99
	AQ(A)A	-	−0.73
	QCH2A	-	−0.71
(d)	AN(A)A	-	−1.68
	OACH2A	+	0.99
	ACH2CH2A	+	0.98
(e)	7 M ring	+	1.78
	C$=C($A)$A	-	−2.14
	NC(O)N	+	1.85
	C=CN	-	−2.03
(f)	CC(C)(C)A	+	1.10
	NAN	-	−8.45
	QHAAACH2A	-	−8.38
	A$A!O > 1	+	1.18
	NH	-	−7.78

[a] Listed are the most important features for the prediction of the compounds shown in Figure 11. Features that are present can be mapped, whereas the influence of features that are absent can only be inferred from the prediction visualization. The odds ratios are reported for feature presence or absence, i.e., $\mathrm{OR}_d = P(x_d = 1 \mid y = 1)/P(x_d = 1 \mid y = 0)$ for present and $\mathrm{OR}_d = P(x_d = 0 \mid y = 1)/P(x_d = 0 \mid y = 0)$ for absent features.

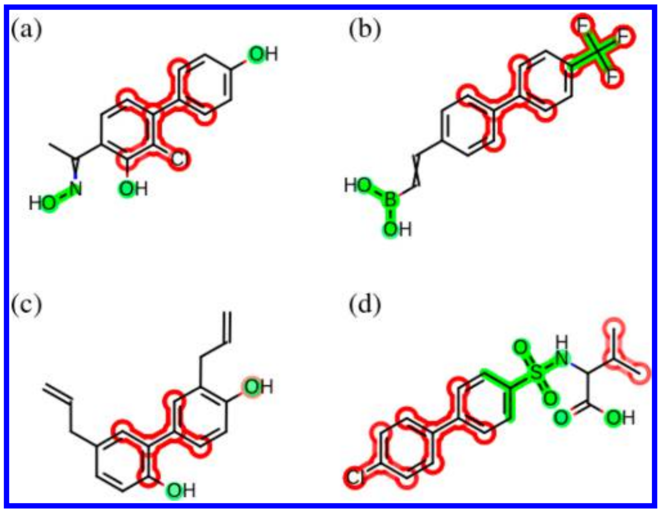


Figure 12. Feature mapping of selected CAI compounds. Features with at least 25% log odds ratio compared to the maximum absolute log OR are back-projected onto four selected CAI compounds including a (a) true positive, (b) true positive, (c) true negative, and (d) false positive. Color shading indicates the magnitude of the log odd ratios.

(−2.13) and the nonset (−1.68) features being negative). Hence, in this case, the influence of set and nonset features is of comparable magnitude.

■ CONCLUDING REMARKS

In this work, a visualization approach for Bayesian classification models and their predictions has been introduced. Naïve Bayesian classifiers are widely used in chemoinformatics. Although Bayesian classification is methodologically less complex than other machine learning approaches such as support vector machines or neural networks, analyzing classification models and rationalizing their performance are far from being trivial tasks. Features that determine the performance of Bayesian models and their potential interplay are difficult to identify, especially if classification proceeds in high-dimensional reference spaces, which is usually the case. In a few instances in which it has thus far been attempted to rationalize the performance of machine learning models, statistical considerations have been applied, for example, in the context of feature selection. Others have used visualization schemes in terms of feature mapping, where they could by design only account for present but not absent features. We have designed a new graphical analysis scheme for "model anatomy" and demonstrated the utility of model visualization and prediction visualization to better understand how Bayesian classification models work. Exemplary compound data sets of different composition and structural heterogeneity, with known or unknown SAR determinants, were used to build Bayesian classification models for activity prediction. On the basis of graphical analysis, we have been able to determine that classification models respond differently to structural characteristics of these compound sets and that feature absence and deselection of inactive compounds often contributes as much (or even more) to prediction accuracy as feature presence and preferential selection of active compounds. The identification of signature features and/or cumulative feature effects play comparably important roles for global model performance and individual predictions. Graphical analysis of the CAI model and representative predictions has demonstrated how the visualization approach introduced herein helps to rationalize model performance and focus on key features. For the more complex data sets MAPK14 (containing kinase inhibitors that are structurally very similar to inactives) and CGRPR (with high structural heterogeneity among actives), the visualizations have enabled us to better understand why classification models reach reasonable to good predictive performance even in these rather difficult cases. Here, our findings highlight the role of compound deselection and cumulative feature effects referred to above. Taken together, our results suggest that model visualization, as introduced herein, should aid in the rationalization and further refinement of Bayesian classification methods. The visualization approach should also be adaptable for other supervised machine learning methods and help reduce their often cited "black box" character. A prototypic Python implementation of our visualization methodology is made freely available via the public Zenodo platform.[44] This implementation should provide a basis for further exploration and extension of our visualization approach.

■ AUTHOR INFORMATION

Corresponding Author

*Phone: +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 authors thank Ye Hu and Norbert Furtmann for help with data collection and feature analysis. The use of OpenEye's OEChem and OEDepict Toolkit was made possible by their free academic licensing program.

REFERENCES

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

(2) Vogt, M.; Bajorath, J. Chemoinformatics: A View of the Field and Current Trends in Method Development. *Bioorg. Med. Chem.* **2012**, *20*, 5317−5323.

(3) Geppert, H.; Vogt, M.; Bajorath, J. Current Trends in Ligand-based Virtual Screening: Molecular Representations, Data Mining Methods, New Application Areas, and Performance Evaluation. *J. Chem. Inf. Model.* **2010**, *50*, 205−216.

(4) Ballester, P. J.; Mitchell, J. B. O. A Machine Learning Approach to Predicting Protein-Ligand Binding Affinity with Applications to Molecular Docking. *Bioinformatics* **2010**, *26*, 1169−1175.

(5) Frank, E.; Bouckaert, R. R. Naive Bayes for Text Classification with Unbalanced Classes. *Proc. 10th European Conf. on Principle and Practice of Knowledge Discovery in Databases* **2006**, 503−510.

(6) Heikamp, K.; Hu, X.; Yan, A.; Bajorath, J. Prediction of Activity Cliffs Using Support Vector Machines. *J. Chem. Inf. Model.* **2012**, *52*, 2354−2365.

(7) Hert, J.; Willett, P.; Wilton, D. J. New Methods for Ligand-Based Virtual Screening: Use of Data Fusion and Machine Learning To Enhance the Effectiveness of Similarity Searching. *J. Chem. Inf. Model.* **2006**, *46*, 462−470.

(8) Prince, S. J. D. *Computer Vision: Models, Learning, and Inference*; Cambridge University Press: 2012.

(9) Wawer, M.; Lounkine, E.; Wassermann, A. M.; Bajorath, J. Data Structures and Computational Tools for the Extraction of SAR Information from Large Compound Sets. *Drug Discovery Today* **2010**, *15*, 630−639.

(10) Wassermann, A. M.; Wawer, M.; Bajorath, J. Activity Landscape Representations for Structure-Activity Relationship Analysis. *J. Med. Chem.* **2010**, *53*, 8209−8223.

(11) Stumpfe, D.; Bajorath, J. Methods for SAR Visualization. *RSC Adv.* **2012**, *2*, 369−378.

(12) Whitesides, G. M.; Krishnamurthy, V. M. Designing Ligands to Bind Proteins. *Q. Rev. Biophys.* **2005**, *38*, 385−395.

(13) Stumpfe, D.; Bajorath, J. Similarity Searching. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2011**, *1*, 260−282.

(14) Glick, M.; Klon, A. E.; Acklin, P.; Davies, J. W. Enrichment of Extremely Noisy High-Throughput Screening Data Using a Naïve Bayes Classifier. *J. Biomol. Screening* **2004**, *9*, 32−36.

(15) Glick, M.; Jenkins, J. L.; Nettles, J. H.; Hitchings, H.; Davies, J. W. Enrichment of High-Throughput Screening Data with Increasing Levels of Noise Using Support Vector Machines, Recursive Partitioning, and Laplacian-Modified Naive Bayesian Classifiers. *J. Chem. Inf. Model.* **2006**, *46*, 193−200.

(16) Glick, M.; Davies, J. W.; Jenkins, J. L. Prediction of Biological Targets for Compounds Using Multiple-Category Bayesian Models Trained on Chemogenomics Databases. *J. Chem. Inf. Model.* **2006**, *46*, 1124−1133.

(17) Xia, X.; Maliski, E. G.; Gallant, P.; Rogers, D. Classification of Kinase Inhibitors Using a Bayesian Model. *J. Med. Chem.* **2004**, *47*, 4463−4470.

(18) Rogers, D.; Brown, R. D.; Hahn, M. Using Extended-Connectivity Fingerprints with Laplacian-Modified Bayesian Analysis in High-Throughput Screening Follow-Up. *J. Biomol. Screening* **2005**, *10*, 682−686.

(19) Sutherland, J. J.; Higgs, R. E.; Watson, I.; Vieth, M. Chemical Fragments as Foundations for Understanding Target Space and Activity Prediction. *J. Med. Chem.* **2008**, *51*, 2689−2700.

(20) Wassermann, A. M.; Kutchukian, P. S.; Lounkine, E.; Luethi, T.; Hamon, J.; Bocker, M. T.; Malik, H. A.; Cowan-Jacob, S. W.; Glick, M. Efficient Search of Chemical Space: Navigating from Fragments to Structurally Diverse Chemotypes. *J. Med. Chem.* **2013**, *56*, 8879−8891.

(21) Klon, A. E.; Glick, M.; Davies, J. W. Application of Machine Learning To Improve the Results of High-Throughput Docking Against the HIV-1 Protease. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2216−2224.

(22) Klon, A. E.; Glick, M.; Thoma, M.; Acklin, P.; Davies, J. W. Finding More Needles in the Haystack: A Simple and Efficient Method for Improving High-Throughput Docking Results. *J. Med. Chem.* **2004**, *47*, 2743−2749.

(23) Klon, A. E.; Glick, M.; Davies, J. W. Combination of a Naive Bayes Classifier with Consensus Scoring Improves Enrichment of High-Throughput Docking Results. *J. Med. Chem.* **2004**, *47*, 4356−4359.

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

(25) Sun, H. A Naïve Bayes Classifier for Prediction of Multidrug Resistance Reversal Activity on the Basis of Atom Typing. *J. Med. Chem.* **2005**, *48*, 4031−4039.

(26) Bender, A.; Mussa, H. Y.; Gill, G. S.; Glen, R. C. Molecular Surface Point Environments for Virtual Screening and the Elucidation of Binding Patterns (MOLPRINT 3D). *J. Med. Chem.* **2004**, *47*, 6569−6583.

(27) Nigsch, F.; Bender, A.; Jenkins, J. L.; Mitchell, J. B. O. Ligand-Target Prediction Using Winnow and Naive Bayesian Algorithms and the Implications of Overall Performance Statistics. *J. Chem. Inf. Model.* **2008**, *48*, 2313−2325.

(28) Rosenbaum, L.; Hinselmann, G.; Jahn, A.; Zell, A. Interpreting Linear Support Vector Machine Models with Heat Map Molecule Coloring. *J. Cheminf.* **2011**, 3, 11.

(29) Riniker, S.; Landrum, G. A. Similarity Maps - a Visualization Strategy for Molecular Fingerprints and Machine-Learning Methods. *J. Cheminf.* **2013**, *5*, 43.

(30) Alpaydin, E. *Introduction to Machine Learning*, 2nd ed.; MIT Press: Cambridge, USA, 2010.

(31) Duda, R. O.; Hart, P. E.; Stork, D. G. *Pattern Classification*, 2nd ed.; Wiley-Interscience: New York, 2000.

(32) Zhang, H. The Optimality of Naïve Bayes. *Proc. 17th Int. Florida Artific. Intell. Res. Soc. Conf.* **2004**, 562−567.

(33) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery. *Nucleic Acids Res.* **2012**, *40*, D1100−D1107.

(34) Dimova, D.; Iyer, P.; Vogt, M.; Totzke, F.; Kubbutat, M. H. G.; Schächtele, C.; Laufer, S.; Bajorath, J. Assessing the Target Differentiation Potential of Imidazole-Based Protein Kinase Inhibitors. *J. Med. Chem.* **2012**, *55*, 11067−11071.

(35) ProQinase Free Choice Biochemical Kinase Assays. http://www.proqinase.com/ (accessed Oct 15, 2013).

(36) 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.

(37) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887−2893.

(38) Xu, Y.-J.; Johnson, M. Algorithm for Naming Molecular Equivalence Classes Represented by Labeled Pseudographs. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 181−185.

(39) *MACCS Structural keys*; Accelrys: San Diego, CA, 2011.

(40) *OEChem TK version 2.0.0*; OpenEye Scientific Software: Santa Fe, NM. http://www.eyesopen.com (accessed July 5, 2014).

(41) *RDKit: Open-source cheminformatics*. http://www.rdkit.org (accessed July 5, 2014).

(42) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.

(43) *OEDepict TK version 2.0.0*; OpenEye Scientific Software: Santa Fe, NM. http://www.eyesopen.com (accessed July 5, 2014).

(44) Balfer, J.; Bajorath, J. Visualization and Graphical Interpretation of Bayesian Compound Classification Models. http://dx.doi.org/10.5281/zenodo.11371.