Journal of
CHEMICAL INFORMATION
AND MODELING

Similarity Boosted Quantitative Structure–Activity Relationship—A Systematic Study of Enhancing Structural Descriptors by Molecular Similarity

Tobias Girschick,[†] Pedro R. Almeida,[‡] Stefan Kramer,[*,§] and Jonna Stålring[⊥]

[†]Technische Universität München, Institut für Informatik/I12, Boltzmannstrasse 3, 85748 Garching b. München, Germany

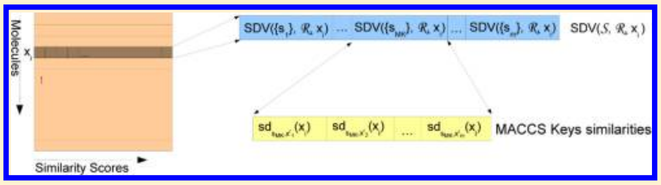
[‡]EngInMotion Ltd, Avenida Infante D. Henrique, n. 145, 3510-070 Viseu, Portugal

[§]Institut für Informatik, Johannes Gutenberg-Universität Mainz, Staudingerweg 9, 55128 Mainz, Germany

[⊥]Computational Toxicology, Global Safety Assessment, AstraZeneca R&D, Pepparedsleden 1, 43183 Mölndal, Sweden

S *Supporting Information*

ABSTRACT: The concept of molecular similarity is one of the most central in the fields of predictive toxicology and quantitative structure–activity relationship (QSAR) research. Many toxicological responses result from a multimechanistic process and, consequently, structural diversity among the active compounds is likely. Combining this knowledge, we introduce similarity boosted QSAR modeling, where we calculate molecular descriptors using similarities with respect to representative reference compounds to aid a statistical learning algorithm in distinguishing between different structural classes. We present three approaches for the selection of reference compounds, one by literature search and two by clustering. Our experimental evaluation on seven publicly available data sets shows that the similarity descriptors used on their own perform quite well compared to structural descriptors. We show that the combination of similarity and structural descriptors enhances the performance and that a simple stacking approach is able to use the complementary information encoded by the different descriptor sets to further improve predictive results. All software necessary for our experiments is available within the cheminformatics software framework AZOrange.


Molecules
x_i
Similarity Scores
SDV({s_1}, R, x_i) ... SDV({s_NN}, R, x_i) ... SDV({s_m}, R, x_i)
SDV(S, R, x_i)
sd(x_i) sd(x_i) ... sd(x_i)
MACCS Keys similarities

INTRODUCTION

Many applications and problem settings in cheminformatics rely on the concept of similarity of small molecules. Examples are search functionalities like substructure search, learning methods like *k*-nearest neighbor as well as variants of virtual screening or clustering. This makes molecular similarity one of the most central concepts in cheminformatics.[1]

One particular application of similarities is their utilization as molecular descriptors. Using similarities of molecular graphs to encode the input space for building (quantitative) structure–activity relationships ((Q)SARs) has been in the air—in one way or another—for some time. Cuadrado et al.[2] present an approach to QSAR based on similarity used as descriptors. They predict a set of 31 steroids using partial least squares (PLS) regression as the learner and an index called Approximate Similarity (AS) as the descriptor space. An all-against-all AS matrix is used as the descriptor set. As this index is based on maximum common subgraph (MCS) calculations, an application to larger data sets is impractical due to immense computational cost raised by the NP-hardness of the problem.[3] Another study that makes use of similarities in the descriptor space has been done by Richter et al.[4] The authors show that they can significantly improve the regression mean absolute error for growth inhibition prediction on NCI DTP human tumor cell line screening data by using background knowledge. The background knowledge in this case are structures and a mode of action grouping of standard anticancer agents (ACAs). This information is encoded and added to the description of the molecules in terms of similarities of the training structures with respect to those ACA reference structures or to the groups of modes of action.

From a machine learning perspective, an elegant example of using similarities in the descriptor space is the concept of the empirical kernel map.[5,6] With the empirical kernel map, any similarity function can be transformed into a valid kernel. To achieve this, a similarity vector is used to represent an instance with respect to the training set. The kernel itself is then defined as the dot product of two similarity vectors. Despite these efforts, many open questions remain, e.g. which similarity measure to use, which reference molecules for the similarity calculations to use, or what the added value of the similarity information is.

In this study we introduce similarity boosted QSAR modeling using chemical similarity scores as descriptors. Our basic idea is to include knowledge about the similarity with respect to a set of reference structures as descriptors. The motivation behind this is that many toxicity responses result from multimechanistic processes and, consequently, there can be structural diversity among the active compounds. The

Received: April 11, 2012
Published: March 14, 2013

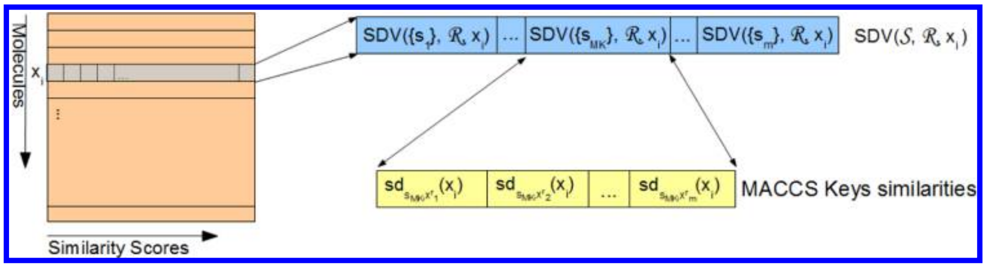

Molecules
x_i
Similarity Scores
SDV({s_1}, $\mathcal{R}$, x_i)
...
SDV({s_{MK}}, $\mathcal{R}$, x_i)
...
SDV({s_m}, $\mathcal{R}$, x_i)
SDV($\mathcal{S}$, $\mathcal{R}$, x_i)
sd$_{s_{MK},x^r_1}$(x_i)
sd$_{s_{MK},x^r_2}$(x_i)
...
sd$_{s_{MK},x^r_m}$(x_i)
MACCS Keys similarities

Figure 1. Schematic depiction of similarity descriptor vector composition for molecule x_i. MACCS keys fingerprint similarity is abbreviated with s_{MK}. $\mathcal{R}$ is the set of reference compounds.

derived similarity scores with respect to representative active compounds aid a statistical learning algorithm in recognizing the various activity classes. Furthermore, extensive sets of reference compounds can be used to span the chemical space in the form of a structural representation, thereby positioning a molecule in it. Extended reference sets make the approach conceptually similar in the structural domain to the ChemGPS[7] method, where the principal components of the physicochemical properties of a set of reference compounds are used as descriptors. Building upon the work by Richter et al., we perform a systematic study assessing the usefulness of various similarity descriptors over a range of public data sets. We use different similarity measures for small molecules alone or in combination, we experiment with three variants to collect or compute reference molecules for the similarity calculations (one based on literature search and two based on clustering) and we combine the similarity descriptors with different sets of structural descriptors. The latter experiment aims to show that our similarity descriptors encode information complementary to that encoded by state-of-the-art structural descriptors, enhancing predictive performance when used in a classification setting.

The remainder of the paper is structured as follows: In the next section we give a stepwise explanation of technical details of the similarity descriptors as well as the experimental setup. This is followed by an overview of the experimental results and a discussion before we conclude.

■ MATERIALS AND METHODS

In this section we describe how the similarity descriptors are built, as well as the data sets and setup we use in our experimental evaluation. All developed methods are available within the open source cheminformatics package AZOrange.[8]

Similarity Descriptors. Throughout the rest of this paper, molecules will be denoted with the letter x and, if necessary, an index to distinguish between different molecules. To compile a *similarity descriptor vector* (*SDV*) for a molecule x, we use two building blocks: First of all, we need a set of similarity functions $\{s_j(x_a, x_b)\}_{j=1}^{l} \in \mathcal{I}$, with $l = |\mathcal{I}|$ that return real-valued measures of similarity given two molecules x_a and x_b. Second, we need a set of reference compounds $\mathcal{R}$ with respect to which we want to calculate the similarities of our training and test compounds. Molecules used as reference will be denoted x^r. Throughout this article the terms *molecule* and *compound* will be used synonymously, as well as reference molecule and reference compound. More formally, we let $\{x_k^r\}_{k=1}^{m} \in \mathcal{R}$, with $m = |\mathcal{R}|$, denote a set of reference molecules. Given a set of symmetric molecule similarity measures $\mathcal{I}$, we define the *similarity descriptor set* SD($\mathcal{I}$, $\mathcal{R}$) as

$$\text{SD}(\mathcal{I}, \mathcal{R}) = \{sd_{s_1,x_1^r}(\cdot), ..., sd_{s_1,x_m^r}(\cdot), ..., sd_{s_l,x_1^r}(\cdot), ..., sd_{s_l,x_m^r}(\cdot)\} \tag{1}$$

where $sd_{s_j,x_k^r}(\cdot)$, $s_j \in \mathcal{I}$, $x_k^r \in \mathcal{R}$ denotes a similarity descriptor, i.e. a function returning the similarity s_j between a reference molecule x_k^r and its argument. Correspondingly, we obtain the similarity descriptor vector SDV of length $l \times m$ for molecule x_i:

$$\text{SDV}(\mathcal{I}, \mathcal{R}, x_i) = (sd_{s_1,x_1^r}(x_i), ..., sd_{s_1,x_m^r}(x_i), ..., sd_{s_l,x_1^r}(x_i), ..., sd_{s_l,x_m^r}(x_i)), \qquad \forall\ s_j \in \mathcal{I}, \forall\ x_k^r \in \mathcal{R} \tag{2}$$

where $sd_{s_j,x_k^r}(x_i)$ is the descriptor value for molecule x_i for descriptor $sd_{s_j,x_k^r}(\cdot)$. A schematic overview of the SDV composition is shown in Figure 1.

Molecular Similarity Measures. The molecular similarity measures used in the experimental evaluation of our approach are built on five molecular fingerprints available in the AZOrange framework: RDKit [http://rdkit.org] MACCS keys, RDKit topological fingerprints, RDKit extended and functional connectivity fingerprints, ECFP and FCFP, respectively, and RDKit Atom Pairs fingerprints. In accordance with the RDKit reference manual recommendations, the similarity between two fingerprints A and B is either calculated with the Tanimoto similarity coefficient[9] (topological fingerprints, MACCS keys):

$$s_{\text{T}} = \frac{c}{a + b - c} \tag{3}$$

or with the related Dice coefficient[10] (ECFP, FCFP, Atom Pairs):

$$s_{\text{D}} = \frac{2c}{a + b} \tag{4}$$

where c is the number of "1" bits common to both fingerprints, a the number of "1" bits in A, and b the number of "1" bits in B. This results in the similarity measures s_{MK}, s_{topo}, s_{ECFP}, s_{FCFP}, and s_{AP}, respectively. In our evaluation experiments we also use a combination of the five fingerprint similarities for building the SDV. Combination in this case means that all similarity measures in $\mathcal{I}$ are used to calculate the similarity of a compound with respect to the reference compounds in $\mathcal{R}$ and thus the length of the similarity descriptor vector will be five times the length when using just one of the similarity measures. The combination can be understood as union of the single similarity descriptors and the respective set of similarity measures will be denoted $\mathcal{I}_{\text{ALL}}$.

Selection of Reference Compounds. We experiment with three variants of how to obtain the set of reference molecules $\mathcal{R}$ for the SDV calculations. The most intuitive way to obtain knowledge about representative active structures for an assay or another biological problem setting is literature search, which is our first variant (R_{LIT}). The descriptors constructed from similarity calculations with respect to the set of reference molecules $\mathcal{R}^{\mathrm{lit}}$ will be denoted SD($\mathcal{I}$, $\mathcal{R}^{\mathrm{lit}}$). The rationale is that we want to use a priori knowledge on different scaffolds or types of actives for the assay at hand. We use one representative for each of those types as reference compound and calculate the similarities of our data set with respect to those reference compounds. A detailed list of the $\mathcal{R}^{\mathrm{lit}}$ reference compounds is given in the Supporting Information. The second and third variant are based on structural clustering.[11,12] The reference compounds $\mathcal{R}^{\mathrm{act}}$ in variant two (variant denoted R_{ACT}; resulting in similarity descriptors SD($\mathcal{I}$, $\mathcal{R}^{\mathrm{act}}$) are cluster representatives from clustering all active compounds of an assay. Extending the number of active reference compounds as compared to R_{LIT}, might account for additional activity classes and hence mechanisms of action not covered by compounds resulting from a manual literature inspection. The clustering algorithm produces overlapping (nondisjoint) and nonexhaustive clusters. The parameters used for the structural clustering procedure are given in Table 1. The thr and minSize parameters are chosen in such a way that the number of clusters is roughly comparable for all seven data sets. We select one compound as cluster representative randomly. Consequently, variant two can be seen as an automated version of variant one, where the different types of actives are found by clustering (although, in case of R_{LIT} the reference compounds do not have to be contained in the assay set). As this second variant uses information about the class of a compound (only actives are clustered), extra care has to be taken during validation. Hence, to ensure a strict validation process the clustering is repeated for each fold, clustering only the actives contained in the training set. In variant three (R_{DB}), we cluster a subset of 300 000 compounds [the subset has been generated by random sampling from the nearly five million commercially available small molecules in the ChemDB] of the ChemDB[13] to obtain the reference compounds $\mathcal{R}^{\mathrm{db}}$ (with resulting descriptors SD($\mathcal{I}$, $\mathcal{R}^{\mathrm{db}}$)). Here, the database subset represents the available chemical space and we want to position the molecules in our data set relative to representative compounds from the chemical space. This makes the third variant a more generic approach to the problem than variants one and two. As the number of clusters is relatively high due to the size of the clustered database, we set the minimum size of a cluster to provide a reference compound to 1500 resulting in 201 reference compounds in $\mathcal{R}^{\mathrm{db}}$.

Table 1. Parameters of the Structural Clustering Algorithm[a]

data set	AID	thr	minSize	$\mathrm{nClusters}_{\mathrm{RACT}}$
hERG	1511	0.6	5	126
AhR	2796	0.7	10	75
ER	639	0.5	5	49
SRC-1	631	0.4	5	53
THR	1479	0.4	5	50
KCNQ2	2156	0.7	5	45
M1	677	0.4	5	56

[a] thr is the similarity threshold for a compound to be added to a cluster. minSize is the minimum size (number of compounds) of a cluster to provide a reference structure, while $\mathrm{nClusters}_{\mathrm{RACT}}$ is the number of clusters resulting from using the R_{ACT} method (mean value from the 100-times repeated hold-out experiments). $\mathrm{nClusters}_{\mathrm{RACT}}$ is equivalent to the size of $\mathcal{R}^{\mathrm{act}}$.

Core Descriptors. One of the goals of this work is to show potential improvement with respect to state-of-the-art structural representations. Consequently, we not only evaluate how our similarity descriptors perform, but we also assess if adding our similarity descriptors to a set of core descriptors improves the performance of a prediction model. We use two sets of structural core descriptors: Backbone Refinement Class (BBRC) descriptors[14] and Extended-Connectivity Fingerprints (ECFP).[15] Please note that the ECFP descriptors used as core descriptors are different than the ones used in the similarity descriptors, although they are compiled with the same algorithm. The difference is in the parametrization and the way the fingerprint bits are used. When used as core descriptors, their descriptor values are used directly as input to the learning algorithms; when used as fingerprint for the similarity descriptors, they are used to calculate the similarity with respect to $\mathcal{R}$ and build the SDV. The algorithm compiling BBRC descriptors mines for frequently occurring class-correlated substructural features of molecules. Class-correlated means that the extracted substructural features not only have to occur frequently but also have to show a significant correlation with the active class. Here, the significance is estimated with a chi-squared p-value lower bound. For the smaller data sets (n < 5000 instances), we use an absolute minimum support parameter of minsup = 150; for the larger ones (n > 5000 instances), we use minsup = 500. As chi-squared significance parameter for the class correlation, we use the default value of ChisqSig = 0.95. Remaining parameters are left at default values if not mentioned otherwise. When considering the results for predictions based on BBRC descriptors later in the paper, please keep in mind that the minsup and ChisqSig parameters are not optimized in any way and performance improvements can be gained by doing so. The algorithm used to compile the BBRC descriptors was integrated into the AZOrange software and is available via the getStructuralDesc.py module. The

Table 2. Summary of the Used PubChem Assay Data Sets and the Number of Descriptors in Each Descriptor Set[a]

data set	n	BBRC	ECFP_{r1}	$\lvert\mathrm{SD}(\mathcal{I}_{\mathrm{ALL}}, \mathcal{R}^{\mathrm{lit}})\rvert$	$\lvert\mathrm{SD}(\mathcal{I}_{\mathrm{ALL}}, \mathcal{R}^{\mathrm{act}})\rvert$	$\lvert\mathrm{SD}(\mathcal{I}_{\mathrm{ALL}}, \mathcal{R}^{\mathrm{db}})\rvert$
hERG	3104	142	1526	30	630	1005
AhR	15980	257	1989[b]	60	375	1005
ER	2302	147	1160	35	245	1005
SRC-1	1622	117	1158	35	265	1005
THR	1632	88	1442	25	250	1005
KCNQ2	6814	172	2119	25	225	1005
M1	1446	28	1123	70	280	1005

[a] The number of examples n comprises 50% actives and 50% inactives. [b] No mean value, only one training fold.

Extended-Connectivity Fingerprint descriptors are circular fingerprint descriptors that use as input information not only the atom and bond type, but the six atom numbering independent Daylight atomic invariants[16] to encode atoms: the number of immediate heavy atom neighbors, the valence minus the number of hydrogens, the atomic number, the atomic mass, the atomic charge, the number of attached hydrogens, plus a seventh invariant added by Rogers et al.:[15] whether the atom is contained in at least one ring. The ECFP descriptor values were calculated with the RDKit functionality of AZOrange. We use default settings and set the radius parameter to $r = 1$ (ECFP$_{r1}$). Table 2 displays the number of compounds in each data set and the dimensionality of the calculated descriptor vectors. The size of the BBRC, ECFP$_{r1}$, and SD($\mathcal{I}$, $\mathcal{R}^{\text{act}}$) are mean values of the 100 hold-out training sets, as these feature sets are data-dependent.

Data Sets. To evaluate the performance of our similarity descriptors we gathered seven data sets from the public database PubChem BioAssays,[17] related to toxicologically relevant end points. The PubChem variable "PUBCHEM_ACTIVITY_OUTCOME" was used as the categorical response variable, and due to the computational requirements, the study was restricted to binary classifiers. To avoid problems related to unbalanced data sets, which are considered outside the scope of this study, inactive compounds were randomly deselected from the PubChem data sets to ensure an equal distribution of active and inactive structures. A tabular overview of all data sets including PubChem BioAssay ID (AID), end point, and number of instances n is given in Table 2 (left-hand side). The first data set (hERG; PubChem AID 1511) originates from a primary cell-based high-throughput screening assay for identification of compounds that protect hERG from block by proarrhythmic agents. Its size is 3104 instances (1552 active compounds and 1552 inactives). The second data set (AhR; AID 2796) is compiled from a luminescence-based primary cell-based high throughput screening assay to identify activators of the aryl hydrocarbon receptor. This is the largest data set of our study, with 15 980 compounds. The third data set (ER; AID 639) contains 2302 compounds from a high throughput screening for estrogen receptor-α coactivator binding potentiators. The fourth data set (SRC-1; AID 631) is comprised of 1622 instances from a primary biochemical high throughput screening assay for agonists of the steroid receptor coactivator 1 recruitment by the peroxisome proliferator-activated receptor gamma (PPARgamma). The fifth data set (THR; AID 1479) is compiled from a total fluorescence counter screen for inhibitors of the interaction of thyroid hormone receptor and steroid receptor coregulator 2 (SCR-2). Its size is 1632 compounds. The sixth data set (KCNQ2; AID 2156) consists of 6814 chemicals from a primary cell-based high-throughput screening assay for identification of compounds that inhibit KCNQ2 potassium channels. Finally, the seventh data set (M1; AID 677) results from an antagonist confirmation screen aiming to discover novel allosteric modulators of the M1 muscarinic receptor and it contains 1446 compounds.

Experimental Setup. The experimental evaluation of our approach was done with two validation strategies (cross-validation and hold-out validation), using the two learning algorithms random forest (RF)[18] and support vector machine (SVM)[19] (CvSVM with RBF kernel), as provided in AZOrange. RF and SVM were selected as examples of popular and conceptionally different machine learning methods used by the QSAR community. The more basic experiments comparing single similarity measures with their combination and the experiments comparing the individual reference molecules selection variants were evaluated with a 10-fold cross-validation with the two learning algorithms. The remaining experiments, where we also assess the statistical significance of our findings with respect to data sampling, were conducted with a 100 times repeated hold-out evaluation with a 2:1 training set test set split ratio. The reason for this is that estimating the statistical significance of the difference of two classifiers in this way is easier to establish as compared to cross-validation. For those experiments, only random forest was used due to running time issues. The statistical evaluation was done with a corrected resampled paired t test[20] at a 95% significance level. The main difference to a standard t test is that it takes into account the high type I error the t test produces in conjunction with random subsampling,[21] which is due to the statistical dependence of the samples. In the result section we report prediction accuracy for cross-validation results and mean accuracy values with their respective standard deviations [please note that standard deviations only quantify the scatter among the values and do not allow for any conclusions on the statistical significance of the difference of the means[22]] for hold-out experiments. The prediction accuracy is calculated as follows:

$$\text{accuracy} = \frac{\text{correct predictions}}{\text{overall predictions}} \qquad (5)$$

and thus represents the percentage of correct predictions.

In both validation scenarios (hold-out and cross-validation), an internal 5-fold cross-validation for model hyper-parameter optimization was applied. For random forest, the number of randomly selected descriptors at each node splitting in the constituting decision trees (nActVars) was optimized over all integers from $1/4(n_{\text{desc}})^{1/2}$ to $(1/2)n_{\text{desc}}$ with increments of $1/4(n_{\text{desc}})^{1/2}$, with n_{desc} being the number of descriptors in the training set. For the SVM, the C and the γ parameters were optimized in the ranges $C \in (2^{-5}, 2^{-3}, ..., 2^{15})$ and $\gamma \in (2^{3}, 2^{1}, ..., 2^{-15})$. These optimization intervals are defaults from the AZOrange software framework. As the running times for SVMs with internal cross-validation for parameter optimization are quite excessive for the larger data sets (see Figure 2), we set a time threshold of 21 days [21 days on a single AMD Athlon 64 X2 Dual Core 5200+ machine with 2.6 GHz, 512 KB cache,

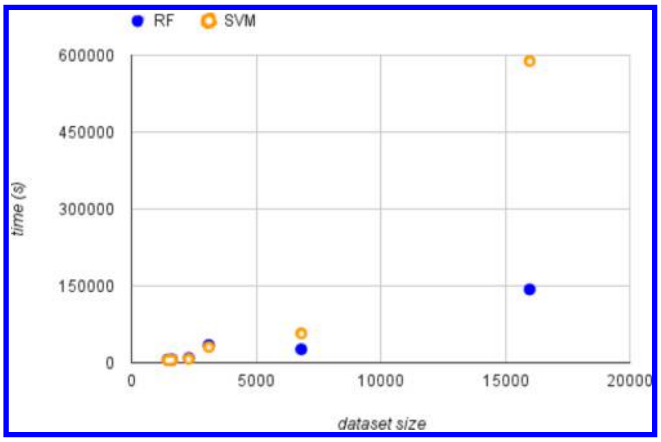


Figure 2. Scatter plot of running times. Shown are mean values of 10-fold cross-validation running times with random forest (RF) and support vector machine (SVM) for the descriptor set SD($\mathcal{I}_{\text{ALL}}$, $\mathcal{R}^{\text{act}}$).

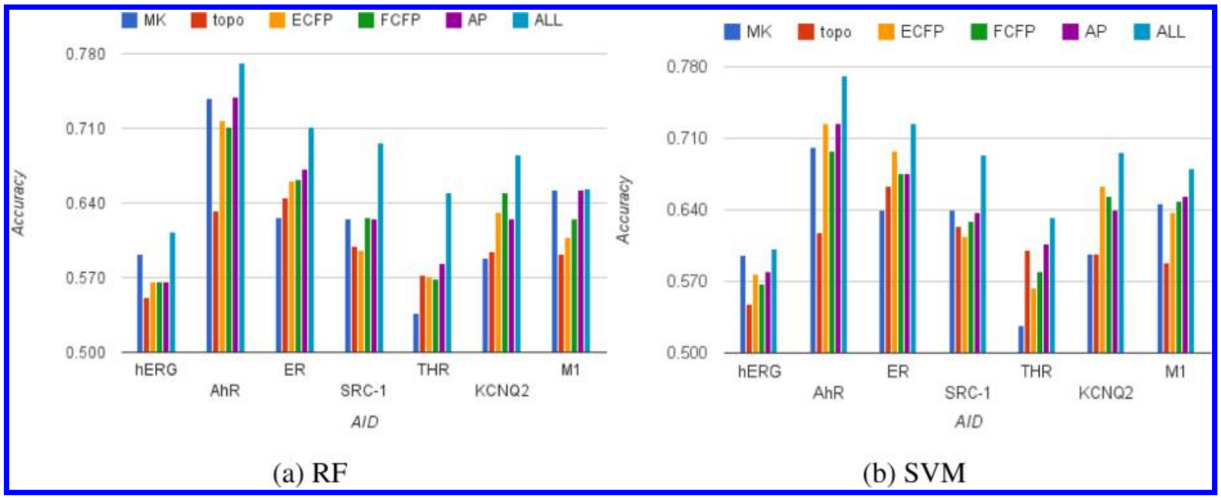


Figure 3. Bar charts of the classification accuracy in a 10-fold cross validation with random forest (RF) and support vector machine (SVM) models based on similarity descriptors with different sets of similarity measures $\mathcal{I}$. In the key of the chart, only the indices of the similarity measure (or set thereof) are given. Consequently, MK corresponds to SD($\{s_{\mathrm{MK}}\}$, $\mathcal{R}^{\mathrm{lit}}$) (analogously for topo, ECFP, FCFP, AP), and ALL to SD($\mathcal{I}_{\mathrm{ALL}}$, $\mathcal{R}^{\mathrm{lit}}$). The set of reference compounds $\mathcal{R}$ is always set to $\mathcal{R}^{\mathrm{lit}}$ in these experiments.

and 4GB DDR2 SDRAM 800] after which experiments are terminated—only results obtained within that time frame are reported in the following. Cells marked with ** in the result tables or figures reflect this time constraint. Tables underlying the result figures and further additional result tables are shown in the Supporting Information.

The experiments are conducted in three consecutive steps: In a first step, we analyze the performance that is achievable with the five similarity metrics used individually. The second experiment series evaluates the three strategies to select the reference molecules for the similarity descriptor calculations, before the combination of structural and similarity descriptors is evaluated in the third experiment.

■ RESULTS

In this section we show and discuss our experimental results in a stepwise manner: First, we analyze the performance of the individual similarity measures s_{MK}, s_{topo}, s_{ECFP}, s_{FCFP}, and s_{AP} and their combination in the set $\mathcal{I}_{\mathrm{ALL}}$. Second, we try to find out which of the three methods to select the reference molecules for similarity calculations—R_{LIT}, R_{ACT}, or R_{DB}—works best. In a third step, we assess the performance of combining the structural core descriptors with our similarity descriptors. This is done twice, in a simple approach of only pooling together the two types of descriptor sets and in an ensemble method approach.

Similarity Descriptors. In the first experiment we compare the performance with solely one fingerprint type in the similarity descriptor vector to the performance achieved when including all five types ($\mathcal{I} = \mathcal{I}_{\mathrm{ALL}}$) as displayed in Figure 1. We select the literature review variant for providing the set of reference molecules ($\mathcal{R}^{\mathrm{lit}}$), while there is no reason to expect different relative results with any of the clustering methods. Results compiled with the random forest learner are shown in Figure 3a and with SVMs in Figure 3b. Looking at the accuracies, we can say that pooling the five basic similarity measures always gives an improvement in predictive performance (the SD($\mathcal{I}_{\mathrm{ALL}}$, $\mathcal{R}^{\mathrm{lit}}$) descriptors are always the rightmost bar in a block). For random forest, we show that this finding is statistically significant in all cases (see Supporting Information Table S11). The conclusion that can be drawn is that the five similarity measures applied in this study together outperform the results achieved individually. In the following, we consequently only consider the similarity descriptors with $\mathcal{I} = \mathcal{I}_{\mathrm{ALL}}$.

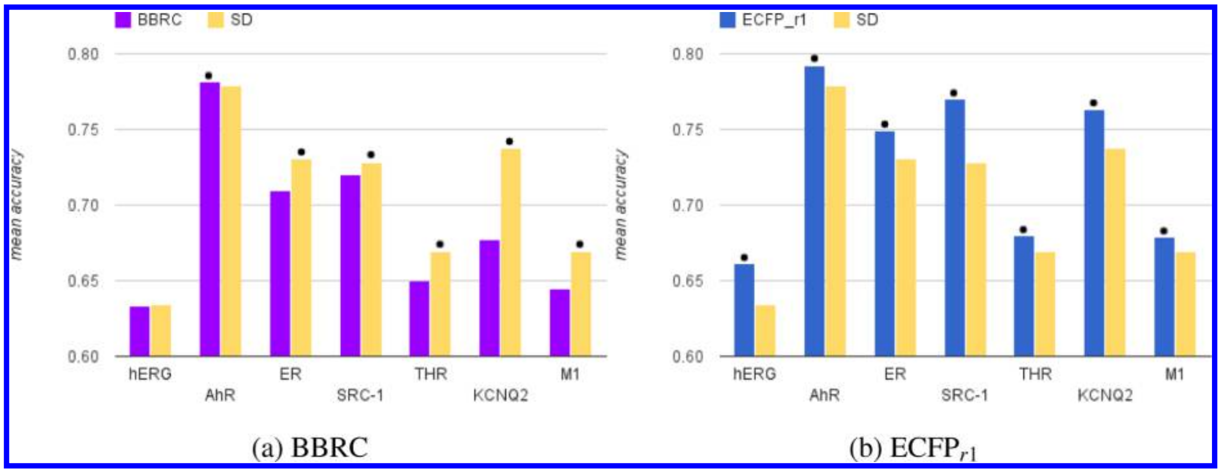
Selection of Reference Molecules. The next experiment compares the descriptor sets based on the different strategies for selecting the reference compounds to calculate the similarities resulting in three descriptor sets: based on manual selection from background knowledge (SD($\mathcal{I}_{\mathrm{ALL}}$, $\mathcal{R}^{\mathrm{lit}}$)), based on clustering the assay actives (SD($\mathcal{I}_{\mathrm{ALL}}$, $\mathcal{R}^{\mathrm{act}}$)) and based on clustering a database representing the chemical space (SD($\mathcal{I}_{\mathrm{ALL}}$,$\mathcal{R}^{\mathrm{db}}$)). A tabular overview of the cross-validation results is given in Table 3.

Table 3. Random Forest (RF) and SVM 10-fold Cross-Validation Prediction Accuracies Using Similarity Descriptors Only[a]

	RF			SVM		
data set	$\mathcal{R}^{\mathrm{lit}}$	$\mathcal{R}^{\mathrm{act}}$	$\mathcal{R}^{\mathrm{db}}$	$\mathcal{R}^{\mathrm{lit}}$	$\mathcal{R}^{\mathrm{act}}$	$\mathcal{R}^{\mathrm{db}}$
hERG	0.613	**0.635**	0.630	0.602	**0.660**	0.653
AhR	0.772	**0.781**	0.774	0.772	**0.807**	**
ER	0.711	0.734	**0.738**	0.725	**0.747**	0.730
SRC-1	0.696	0.730	**0.743**	0.694	**0.761**	0.743
THR	0.650	**0.673**	0.644	0.633	**0.691**	0.655
KCNQ2	0.686	**0.742**	0.732	0.697	**0.777**	0.768
M1	0.653	**0.694**	0.678	0.680	**0.693**	0.690
wins	0	5	2	0	6	0

[a] The three descriptor sets based on similarity with respect to reference molecules are shown. The best descriptor set per learning algorithm is marked in bold print. Column headers only give the set of reference compounds, $\mathcal{I}$ is always $\mathcal{I}_{\mathrm{ALL}}$.

The manual selection variant R_{LIT} is outperformed by the two clustering variants with both learning algorithms. In addition, the clustering methods do not require any manual work searching the literature for scaffold representatives or a priori mechanistic understanding. Comparing the two clustering variants we see that the SD($\mathcal{I}_{\mathrm{ALL}}$, $\mathcal{R}^{\mathrm{act}}$) descriptors outperform the SD($\mathcal{I}_{\mathrm{ALL}}$, $\mathcal{R}^{\mathrm{db}}$) descriptors in five of seven



(a) BBRC (b) ECFP$_{r1}$

Figure 4. Bar charts showing the predictive accuracies for BBRC vs SD($\mathcal{I}_{ALL}$, $\mathcal{R}^{act}$) and ECFP$_{r1}$ vs SD($\mathcal{I}_{ALL}$, $\mathcal{R}^{act}$). The similarity descriptors are abbreviated SD in the key of the chart. Bars representing results that are significantly better than their corresponding neighbor are marked with a black dot on top.

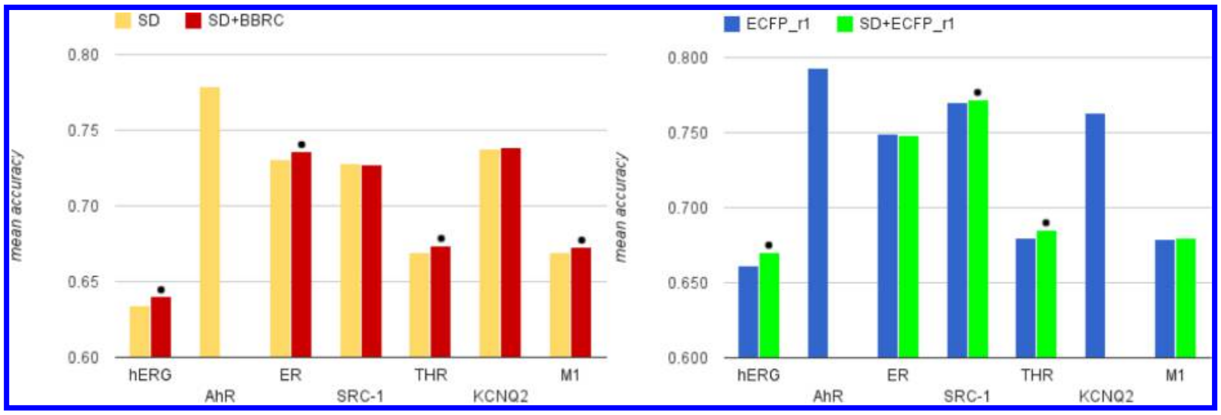


Figure 5. Bar charts showing the predictive accuracies for SD($\mathcal{I}_{ALL}$, $\mathcal{R}^{act}$) (SD in the key) and the combination with BBRC and ECFP$_{r1}$. Bars representing results that are significantly better than the corresponding SD($\mathcal{I}_{ALL}$, $\mathcal{R}^{act}$) or ECFP$_{r1}$ bar are marked with a black dot. SD + BBRC and SD + ECFP$_{r1}$ denote the classifiers built on the combined descriptor sets.

Table 4. Statistical Significance Analysis of Improvement When Combining SD($\mathcal{I}_{ALL}$, $\mathcal{R}^{act}$) (SD in the Table Header) with Structural Descriptors[a]

	random forest			
data set	SD	SD + BBRC	ECFP$_{r1}$	SD + ECFP$_{r1}$
hERG	0.634 ± 0.015	0.640 ± 0.011[b]	0.661 ± 0.012	0.670 ± 0.014[b]
AhR	0.779 ± 0.008	** ± **	0.792 ± 0.015	** ± **
ER	0.731 ± 0.014	0.736 ± 0.014[b]	0.749 ± 0.014	0.748 ± 0.016
SRC-1	0.728 ± 0.018	0.727 ± 0.016	0.770 ± 0.016	0.772 ± 0.011[b]
THR	0.669 ± 0.016	0.674 ± 0.011[b]	0.680 ± 0.018	0.685 ± 0.014[b]
KCNQ2	0.738 ± 0.010	0.739 ± 0.009	0.763 ± 0.009	** ± **
M1	0.669 ± 0.017	0.673 ± 0.019[b]	0.679 ± 0.021	0.680 ± 0.017

[a] The null hypothesis is that there is no improvement compared to the SD($\mathcal{I}_{ALL}$, $\mathcal{R}^{act}$) or ECFP$_{r1}$ column. Shown are mean accuracy values ± standard deviations for 100 hold-out runs [please note again that standard deviations only quantify the scatter among the values and do not allow for any conclusions on the statistical significance of the difference of the means[22]]. SD + BBRC and SD + ECFP$_{r1}$ denote the classifiers built on the combined descriptor sets. [b] Statistically significant improvement wrt column SD($\mathcal{I}_{ALL}$, $\mathcal{R}^{act}$) and ECFP$_{r1}$, respectively.

cases using random forest and in all cases using SVM making clustering of training set actives the preferred method for the selection of reference molecules.

The performance of models based solely on similarity descriptors (SD($\mathcal{I}_{ALL}$, $\mathcal{R}^{act}$)) is further compared to models using the two sets of established structural descriptors (core descriptors). Figure 4 and Supporting Information Tables S12 and S13 display the accuracies and a statistical assessment of the differences in a 100 times repeated hold-out validation. We see that SD($\mathcal{I}_{ALL}$, $\mathcal{R}^{act}$) descriptors obtain mean accuracies significantly better than BBRC in five out of seven data sets but also that ECFP$_{r1}$ is better than SD($\mathcal{I}_{ALL}$, $\mathcal{R}^{act}$) in all seven cases. This shows that the similarity descriptors used on their own are competitive to structural descriptors, although they do not outperform the best.

Table 5. Analysis of Classifier Diversity for the AhR and SRC-1 Data Sets[a]

	AhR				SRC-1			
	Q	ρ	DF	χ^2 *p*-value	Q	ρ	DF	χ^2 *p*-value
BBRC and SD	−0.184	−0.084	0.111	$<10^{-3}$	−0.106	−0.050	0.141	0.027
ECFP$_{r1}$ and SD	−0.195	−0.089	0.107	$<10^{-3}$	−0.143	−0.066	0.125	0.023

[a] Given are hold-out mean values for diversity measures of BBRC and SD($\mathcal{I}_{\mathrm{ALL}}$, $\mathcal{R}^{\mathrm{act}}$) based classifiers as well as for ECFP and SD($\mathcal{I}_{\mathrm{ALL}}$, $\mathcal{R}^{\mathrm{act}}$) based classifiers. SD($\mathcal{I}_{\mathrm{ALL}}$, $\mathcal{R}^{\mathrm{act}}$) is abbreviated SD in the table.

Table 6. Prediction Accuracy Results (± Standard Deviations) of the Two Stacking Variants Combining the SD($\mathcal{I}_{\mathrm{ALL}}$, $\mathcal{R}^{\mathrm{act}}$) and ECFP$_{r1}$ Descriptor Set Based Classifiers in Comparison to the ECFP$_{r1}$ and SD($\mathcal{I}_{\mathrm{ALL}}$, $\mathcal{R}^{\mathrm{act}}$) + ECFP$_{r1}$ Based Classifiers[a]

	random forest			
data set	ECFP$_{r1}$	SD + ECFP$_{r1}$	Stacking$_{\mathrm{mean}}$	Stacking$_{\mathrm{mult}}$
AhR	0.792 ± 0.015	** ± **	0.835 ± 0.007[b]	0.809 ± 0.028[b]
SRC-1	0.770 ± 0.016	0.772 ± 0.011	0.833 ± 0.007[b]	0.790 ± 0.025[b]

[a] SD($\mathcal{I}_{\mathrm{ALL}}$, $\mathcal{R}^{\mathrm{act}}$) is abbreviated as SD in the table. [b] Statistically significant improvement column SD + ECFP$_{r1}$ (or ECFP$_{r1}$ where results for the former are not available).

Combining Structural Descriptors and Similarity Descriptors. Our last experiment assesses the complementarity of the similarity descriptors and standard structural descriptors used for QSAR modeling. For this purpose, we add the structural BBRC and ECFP$_{r1}$ descriptors to the SD($\mathcal{I}_{\mathrm{ALL}}$, $\mathcal{R}^{\mathrm{act}}$) similarity descriptors and assess the performance of the combined descriptor sets (the union) to see if the similarity descriptors complement the structural descriptors. Figure 5 and Table 4 show the results for the significance analysis done with random forest. All further analyses are conducted with random forest only for run time reasons. An important finding is that the combinations are always either significantly better than the structural core descriptors alone or on par with them. This suggests that there is information complementary to the structural descriptors encoded in the similarity descriptors.

To substantiate the assertion that the similarity descriptors and the structural descriptors are complementary, we analyze the diversity of the classifiers based on BBRC and SD($\mathcal{I}_{\mathrm{ALL}}$, $\mathcal{R}^{\mathrm{act}}$) descriptors as well as that based on ECFP$_{r1}$ and SD($\mathcal{I}_{\mathrm{ALL}}$, $\mathcal{R}^{\mathrm{act}}$) descriptors. As measures of classifier diversity, we calculate the Yule's Q statistic, the correlation coefficient ρ, and the double-fault measure DF as proposed by Kuncheva and Whitaker.[23] The exact mathematical definitions of the three measures are given in the Supporting Information. In addition, we perform a chi-squared test of independence for the two classifiers. The expectation of Q is zero for statistically independent classifiers, with $Q \in (-1, 1)$. Classifiers that predict the same instances correctly will have values of $Q > 0$ and classifiers committing prediction errors on different instances will result in $Q < 0$. The double-fault measure is the proportion of cases in which both classifiers commit a prediction error and smaller values indicate a higher diversity of the classifiers ($\mathrm{DF} \in (0, 1)$). For this analysis, we arbitrarily select the AhR and the SRC-1 data sets as one of the larger and one of the smaller data sets. The results are given in Table 5. If we consider a significance level of 10^{-3} for the chi-squared test of independence with the null hypothesis being that the events of error of both classifiers (one based on SD($\mathcal{I}_{\mathrm{ALL}}$, $\mathcal{R}^{\mathrm{act}}$) and one based on structural descriptors) are independent the results give an unclear picture. For the AhR data set, we have to reject the null hypothesis, for the SRC-1 we accept it. The Q statistic values for both data sets are slightly negative, as are the correlation coefficients. This indicates that the structural descriptor based models (BBRC or ECFP$_{r1}$) and SD($\mathcal{I}_{\mathrm{ALL}}$, $\mathcal{R}^{\mathrm{act}}$) based models commit prediction errors on different instances, which is also reflected by the double-fault measure. The DF value of 0.11 for the AhR data set (BBRC and SD($\mathcal{I}_{\mathrm{ALL}}$, $\mathcal{R}^{\mathrm{act}}$)) means that the theoretical upper accuracy limit to be achieved with an ensemble method working with BBRC and SD($\mathcal{I}_{\mathrm{ALL}}$, $\mathcal{R}^{\mathrm{act}}$) base classifiers is 88.9%. While our goal is clearly to eliminate all errors in case of conflicting predictions, this can hardly be achieved in practice.

The observation of classifier diversity suggests a more evolved approach (than just using all descriptors at once) to combining the structural descriptors and the similarity descriptors based on so-called ensembles.[24,25] Ensembles are combinations of classifiers aiming for the reduction of the errors committed by the individual classifiers. Empirical and theoretical results[24] have shown that there exists a positive correlation between the accuracy of ensembles and the diversity amongst the constituting base classifiers. Diversity in this case means that the base classifiers commit prediction errors on different instances and consequently can complement each other when combined. We applied a simple variant of the ensemble method stacking[25] in such a way that individual random forest models are learned for the similarity descriptors and for the structural descriptors. As random forest (as well as SVM) can provide class probability estimates, we can use a combining function to get a single class probability from the two input class probabilities. The first combining function we applied is simply the mean of the two probabilities (Stacking$_{\mathrm{mean}}$); the second multiplies the input probabilities (Stacking$_{\mathrm{mult}}$). In the second variant the decision threshold for the result class is shifted from the standard value of 0.5 to 0.25. The results for both combining function variants are given in Table 6. Both stacking variants are able to improve the overall mean prediction accuracy significantly by roughly 4% compared to the best results so far.

To further understand the properties of the similarity descriptors, we analyze the mean sensitivity and specificity values corresponding to the BBRC, ECFP$_{r1}$, SD($\mathcal{I}_{\mathrm{ALL}}$, $\mathcal{R}^{\mathrm{lit}}$), SD($\mathcal{I}_{\mathrm{ALL}}$, $\mathcal{R}^{\mathrm{act}}$), and SD($\mathcal{I}_{\mathrm{ALL}}$, $\mathcal{R}^{\mathrm{db}}$) classifications (see Supporting Information Tables S14−S16). Except for SD($\mathcal{I}_{\mathrm{ALL}}$, $\mathcal{R}^{\mathrm{lit}}$), where we observe a slight advantage for the sensitivity, the negative class seems to be predicted marginally better than the positive class by all descriptor sets for all data sets. However, the small differences between the sensitivity and specificity show that none of the descriptor sets can be identified as biased with respect to any of the classes.

CONCLUSIONS

In this paper, we systematically studied similarity boosted QSAR, using chemical similarity to a finite set of selected reference compounds for QSAR modeling. We derived those references with three variants out of which one is based on literature search and two on automatic structural clustering. The two clustering variants outperformed the literature search-based method in our experiments. We suspect that the relative success of the SD($\mathcal{I}$, $\mathcal{R}^{\text{act}}$) descriptors as compared to the results achieved with the SD($\mathcal{I}$, $\mathcal{R}^{\text{lit}}$) descriptors derived from the limited set of activity classes of $\mathcal{R}^{\text{lit}}$ compounds, representing only a subset of the mechanisms responsible for the activity, while the $\mathcal{R}^{\text{act}}$ reference compounds might cover those activity classes better with a potentially greater structural diversity among the reference compounds. The SD($\mathcal{I}$, $\mathcal{R}^{\text{db}}$) descriptors can be understood as a generic representation of chemical structure, in the spirit of ChemGPS in the physicochemical space, perhaps primarily alternative rather than complementary to the BBRC and ECFP$_{r1}$ representations. Keeping in mind that the parameters for the structural clustering have not been optimized at all, there should still be room for performance improvements. For example, the clustering algorithm can be further optimized by a refined selection of cluster representatives or hierarchical clustering variants.

For all three variants of the similarity descriptors we use a combination of five similarity measures. We show that they are complementary to a certain extent as using them in combination ($\mathcal{I}_{\text{ALL}}$) increases the predictive performance. The similarity descriptors could be further enhanced by adding pharmacophore-based, maximum common substructure (MCS) based, or yet other similarity measures. Especially MCS based similarities could be of particular importance in toxicological modeling, as such responses are often triggered by the presence of a larger fragment, rather than the global properties or small fragments of the compound.

An interesting point of discussion is the information content of the different sets of descriptors. The ECFP$_{r1}$ descriptors use information about the structure, based on circular atom neighborhoods. They also incorporate information on atom properties (atomic invariants). The second structural descriptor set, BBRC, is based on important substructural features. Important in this case means that the features are frequent and correlated with the end point variable.

Because it is theoretically possible to construct infinitely many structural features for structured data, such structural descriptor sets pose the difficult challenge[26,27] of selecting a small number of relevant patterns or features from a larger set. The similarity descriptors on the other hand encode information about the chemical similarity with respect to a set of reference compounds and the similarity itself is based on diverse information: MACCS keys, topological information, ECFP and FCFP circular neighborhoods, and atom pair information. Clustering actives of the training set to define the reference compounds (and also in the case of the R_{LIT} method if the reference compounds are part of the training set) and using similarity with respect to these compounds as descriptors can be interpreted as instance selection. This option to reduce data set redundancy and complexity is also provided by kernel machines like support vector machine that intrinsically performs instance selection, as it uses only the support vectors instead of all instances to discriminate between classes.[19] As the success of kernel methods is documented in particular for structured data like graphs,[28] instance selection appears as a promising alternative to feature selection for such data, either during learning (kernel machines) or, as investigated in this paper, during feature generation (similarity descriptors).

In our experiments using random forest and SVM, we showed that similarity descriptors in similarity boosted QSAR modeling perform quite well compared to established structural descriptor sets. In addition, combining similarity descriptors with structural descriptors can often further enhance the performance, improving the accuracies achievable with solely structural descriptors. This indicates that the similarity descriptors encode information complementary to structural descriptors.

We support this finding by a statistical analysis of the diversity of classifiers based on either structural or similarity descriptors. The analysis shows that the different sets of descriptors commit prediction errors on different instances. We use this information in a simple stacking approach that improves the prediction results further and confirms that the structural and the similarity descriptors encode complementary information.

Finally, all methods are interfaced with the publically available AZOrange software framework.

ASSOCIATED CONTENT

Supporting Information

Lists of reference compounds $\mathcal{R}^{\text{lit}}$ are given in Tables S1−S7. Tables S8−S10 contain the data underlying Figures 2 and 5. Table S11 shows the significance analysis for the single similarity measures with random forest. Tables S12 and S13 contain the data underlying Figure 4. Tables S14−S16 show mean sensitivity and specificity values including their differences for BBRC, ECFP$_{r1}$, SD($\mathcal{I}_{\text{ALL}}$, $\mathcal{R}^{\text{lit}}$), SD($\mathcal{I}_{\text{ALL}}$, $\mathcal{R}^{\text{act}}$), and SD($\mathcal{I}_{\text{ALL}}$, $\mathcal{R}^{\text{db}}$). A section with mathematical formulas for the diversity measures is provided. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: kramer@informatik.uni-mainz.de.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We would like to acknowledge Andreas Maunz for help and feedback on BBRC software integration issues and Ernst Ahlberg Helgee for sharing his download of the PubChem BioActivity data sets. In addition, we recognize the valuable and constructive comments from all members of the Computational Toxicology group at AstraZeneca. T.G. acknowledges the support of the Center of Doctoral Studies in Informatics and its Application (CeDoSIA) of TUM Graduate School.

REFERENCES

(1) Nikolova, N.; Jaworska, J. Approaches to measure chemical similarity–a review. *QSAR Comb. Sci.* **2003**, *22*, 1006–1026.

(2) Cuadrado, M.; Ruiz, I.; Gómez-Nieto, M. A Steroids QSAR Approach Based on Approximate Similarity Measurements. *J. Chem. Inf. Model.* **2006**, *46*, 1678–1686.

(3) Garey, M.; Johnson, D. *Computers and Intractability; A Guide to the Theory of NP-Completeness*; W.H.Freeman and Company: San Francisco, 1990.

(4) Richter, L.; Hechtl, S.; Kramer, S. Leveraging Chemical Background Knowledge for the Prediction of Growth Inhibition. In *Sixth IEEE International Symposium on BioInformatics and BioEngineering (BIBE 2006)*, Arlington, Virginia, USA, Oct 16–18, 2006; pp 319–324.

(5) Tsuda, K. Support vector classifier with asymetric kernel function. In *ESANN 1999, 7th European Symposium on Artificial Neural Networks*, Bruges, Belgium, Apr 21–23, 1999; pp 183–188.

(6) Schölkopf, B.; Tsuda, K.; Vert, J. *Kernel methods in computational biology*; The MIT press: Cambridge, MA, 2004.

(7) Oprea, T.; Gottfries, J. Chemography: the art of navigating in chemical space. *J. Comb. Chem.* **2001**, *3*, 157–166.

(8) Stålring, J.; Carlsson, L.; Almeida, P.; Boyer, S. AZOrange-High performance open source machine learning for QSAR modeling in a graphical programming environment. *J. Cheminf. [Online]* **2011**, 3, 28 DOI: 10.1186/1758-2946-3-28.

(9) Tanimoto, T. *Internal Report*; IBM: Armonk, NY, 1957.

(10) Dice, L. R. Measures of the Amount of Ecologic Association Between Species. *Ecology* **1945**, *26*, 297–302.

(11) Seeland, M.; Girschick, T.; Buchwald, F.; Kramer, S. Online Structural Graph Clustering Using Frequent Subgraph Mining. In *Machine Learning and Knowledge Discovery in Databases, European Conference, ECML PKDD 2010*, Barcelona, Spain, Sep 20–24; Balcázar, J. L., Bonchi, F., Gionis, A., Sebag, M., Eds.; Lecture Notes in Computer Science; Springer: New York, 2010; Vol. *6323*, Part III, pp 213–228.

(12) Seeland, M.; Berger, S.; Stamatakis, A.; Kramer, S. Parallel Structural Graph Clustering. In *Machine Learning and Knowledge Discovery in Databases, European Conference, ECML PKDD 2011*, Athens, Greece, Sep 5–9; Gunopulos, D., Hofmann, T., Malerba, D., Vazirgiannis, M., Eds.; Lecture Notes in Computer Science; Springer: New York, 2011; Vol. *6913*, Part III, pp 256–272.

(13) Chen, J.; Linstead, E.; Swamidass, S.; Wang, D.; Baldi, P. ChemDB update—full-text search and virtual chemical space. *Bioinformatics* **2007**, *23*, 2348–2351.

(14) Maunz, A.; Helma, C.; Kramer, S. Efficient mining for structurally diverse subgraph patterns in large molecular databases. *Machine Learning* **2011**, *83*, 193–218.

(15) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754.

(16) Weininger, D.; Weininger, A.; Weininger, J. SMILES. 2. Algorithm for Generation of Unique SMILES Notation. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 97–101.

(17) Wang, Y.; Bolton, E.; Dracheva, S.; Karapetyan, K.; Shoemaker, B.; Suzek, T.; Wang, J.; Xiao, J.; Zhang, J.; Bryant, S. An overview of the PubChem BioAssay resource. *Nucleic Acids Res.* **2010**, *38*, D255–66.

(18) Breiman, L. Random forests. *Machine Learning* **2001**, *45*, 5–32.

(19) Vapnik, V. *The nature of statistical learning theory*; Springer: New York, 1995.

(20) Nadeau, C.; Bengio, Y. Inference for the generalization Error. *Machine Learning* **2003**, *52*, 239–281.

(21) Dietterich, T. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. *Neural Comput.* **1998**, *10*, 1895–1924.

(22) Cumming, G.; Fidler, F.; Vaux, D. Error bars in experimental biology. *J. Cell Biol.* **2007**, *177*, 7–11.

(23) Kuncheva, L.; Whitaker, C. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning* **2003**, *51*, 181–207.

(24) Breiman, L. Stacked Regressions. *Machine Learning* **1996**, *24*, 49–64.

(25) Wolpert, D. Stacked generalization. *Neural Networks* **1992**, *5*, 241–259.

(26) Rückert, U.; Kramer, S. Optimizing feature sets for structured data. In *Machine Learning: ECML 2007, 18th European Conference on Machine Learning*, Warsaw, Poland, Sept 17–21; Kok, J., Koronacki, J., López de Mántaras, R., Matwin, S., Mladenic, D., Eds.; Lecture Notes in Computer Science; Springer: New York, 2007; Vol. *4701*; pp 716–723.

(27) Vreeken, J.; van Leeuwen, M.; Siebes, A. Krimp: mining itemsets that compress. *Data Mining Knowl. Discov.* **2011**, *23*, 169–214.

(28) Gärtner, T. *Kernels for structured data*; World Scientific Pub Co Inc: Hackensack, NJ, 2009; Vol. 72