Journal of Chemical Information and Modeling

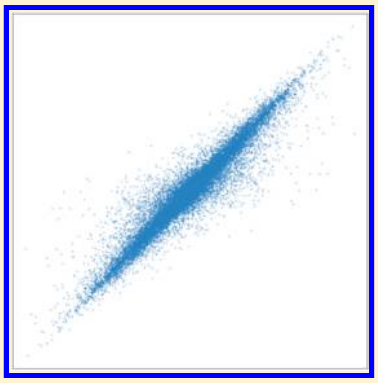
Article

pubs.acs.org/jcim

Prediction of Compound Potency Changes in Matched Molecular Pairs Using Support Vector Regression

Antonio de la Vega de León and Jürgen Bajorath*

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



ABSTRACT: Matched molecular pairs (MMPs) consist of pairs of compounds that are transformed into each other by a substructure exchange. If MMPs are formed by compounds sharing the same biological activity, they encode a potency change. If the potency difference between MMP compounds is very small, the substructure exchange (chemical transformation) encodes a bioisosteric replacement; if the difference is very large, the transformation encodes an activity cliff. For a given compound activity class, MMPs comprehensively capture existing structural relationships and represent a spectrum of potency changes for structurally analogous compounds. We have aimed to predict potency changes encoded by MMPs. This prediction task principally differs from conventional quantitative structure−activity relationship (QSAR) analysis. For the prediction of MMP-associated potency changes, we introduce direction-dependent MMPs and combine MMP analysis with support vector regression (SVR) modeling. Combinations of newly designed kernel functions and fingerprint descriptors are explored. The resulting SVR models yield accurate predictions of MMP-encoded potency changes for many different data sets. Shared key structure context is found to contribute critically to prediction accuracy. SVR models reach higher performance than random forest (RF) and MMP-based averaging calculations carried out as controls. A comparison of SVR with kernel ridge regression indicates that prediction accuracy has largely been a consequence of kernel characteristics rather than SVR optimization details.

■ INTRODUCTION

The prediction of changes in compound potency as a consequence of chemical modifications typically falls into the domain of classical quantitative structure−activity relationship (QSAR) analysis.[1] However, QSAR modeling is usually limited to congeneric compound series[1] and cannot be systematically applied to large and structurally diverse compound data sets. In addition, conventional QSAR predictions are based upon linear regression models. To account for the nonlinearity of many SARs, machine learning approaches such as random forest[2] (RF) and support vector regression[3] (SVR) analysis have gained popularity in recent years. In addition to potency prediction,[4−6] these methods have also been applied, for example, to predict ADME properties[7,8] or toxicology end points.[9−11]

The congeneric compound series constraint of standard QSAR can be addressed by considering alternative structural representations, which can also be combined with nonlinear prediction methods. For example, an attractive opportunity to further extend potency predictions to large and heterogeneous compound data sets is provided by the matched molecular pair (MMP) formalism.[12] An MMP is defined as a pair of compounds that only differ by a structural change at a single site.[12] This modification is accounted for by the exchange of a pair of substructures termed a chemical transformation. MMPs can be systematically generated for a given compound data set, which reveals all possible pairs of structural analogs and comprehensively captures structural relationships present within the set. Then, it can be attempted to predict changes in potency or other chemical properties associated with chemical transformations at the level of compound pairs, rather than series. For example, property value changes in multiple MMPs have been used to predict value changes associated with equivalent transformations.[13,14] However, such MMP-based extrapolations cannot be generalized and have often limited statistical significance.[14]

Accordingly, first attempts have been made to combine MMP analysis with machine learning, for example, by predicting changes in potency and ADME properties using RF calculations[15] or by predicting activity cliffs (i.e., pairs of structurally analogous compounds having a large difference in potency)[16] using support vector machines (SVMs).[17,18] SVM-based activity cliff prediction has distinguished pairs of compounds forming activity cliffs from others,[18] without predicting numerical potency differences. In addition to using SVM models for classification and ranking, prediction of numerical potency (difference) values encoded by MMPs can be attempted via SVR.

Herein, we combine MMP analysis with SVR and systematically predict potency difference values for MMPs using kernel functions of different design. Combinations of different kernel functions and fingerprint descriptors used as molecular fragment representations are explored, and preferred combinations are identified. In calculations on a variety of compound data sets, preferred kernel-fingerprint combinations yield high

Received: July 4, 2014
Published: September 5, 2014

SVR accuracy in predicting numerical potency differences (and reach higher accuracy than RF calculations). The MMP-based SVR methodology introduced herein is generally applicable for numerical property predictions.

■ MATERIALS AND METHODS

Compound Data Sets. From ChEMBL (version 17),[19] 17 compound activity classes including a variety of targets were selected. For each class, all compounds having defined K_i values (with activity relation "=", assay confidence score 9, the highest possible score, and target relationship "D" indicating "direct" relationships) were collected. Compounds having multiple potency values that differed by more than 1 order of magnitude (considering the highest and lowest values) were discarded. For compounds with multiple activity values falling within 1 order of magnitude range, the arithmetic mean was calculated as the final potency annotation. The compound data sets contained between 1200 and 2500 compounds, as reported in Table 1. From these compound data sets, between 5700 and 32 000 direction-dependent MMPs were obtained, as also reported in Table 1.

Table 1. Compound Data Sets[a]

	ID	name	abbreviation	Cpds	MMPs
A	205	carbonic anhydrase II	CA2	1566	8248
B	214	serotonin 1a receptor	5-HT1A	1276	9352
C	217	dopamine D2 receptor	DRD2	1916	17 630
D	218	cannabinoid CB1 receptor	CB1	1673	16 004
E	226	adenosine A1 receptor	ADORA1	2107	24 534
F	228	serotonin transporter	5HTT	1317	9352
G	233	mu opioid receptor	MOR1	1447	15 712
H	234	dopamine D3 receptor	DRD3	1332	9532
I	237	kappa opioid receptor	KOR-1	1302	17 654
J	251	adenosine A2a receptor	ADORA2A	2538	32 086
K	253	cannabinoid CB2 receptor	CB2	1903	18 548
L	256	adenosine A3 receptor	ADORA3	2037	22 316
M	259	melanocortin receptor 4	MC4R	1209	28 274
N	261	carbonic anhydrase I	CA1	1528	7,804
O	264	histamine H3 receptor	H3R	1,849	19 256
P	3371	serotonin 6 receptor	5-HT6	1291	9648
Q	3594	carbonic anhydrase IX	CA9	1220	5756

[a] For each data set, the number of compounds and direction-dependent MMPs is given. In addition, the ChEMBL identifier (ID), the target name, and abbreviation are provided. Data sets are labeled A–Q.

Direction-Dependent Matched Molecular Pairs. Matched molecular pairs were systematically calculated for data set compounds using an in-house Java implementation of the algorithm developed by Hussain and Rea[20] based upon the OEChem Toolkit[21] from OpenEye. Single-, dual-, and triple-cut fragmentation of exocyclic bonds was carried out generating conserved key and variable value fragments stored in an index table.[20] In addition, transformation size restrictions[22] were applied to ensure a meaningful distinction between core structures (key) and substituents (values). Accordingly, a key fragment was required to consist of at least twice the number of non-hydrogen atoms as a value fragment; a value fragment was permitted to contain a maximum of at most 13 non-hydrogen atoms, and the size difference between values of a given key was limited to at most eight non-hydrogen atoms.[22] If two compounds formed several MMPs, the one having the largest key fragment was selected.

For each MMP, the potency difference between the participating compounds was recorded in a direction-dependent manner, as illustrated in Figure 1A. Thus, for each original MMP, two direction-dependent MMPs were generated encoding a potency-decreasing and a potency-increasing transformation (i.e., value fragment 1 → 2 vs 2 → 1).

Molecular Representation. For each direction-dependent MMP, five fingerprint representations were calculated, as also illustrated in Figure 1A, including a fingerprint of the key fragment (KeyFP), fingerprint of the value fragments (V1FP and V2FP, respectively), a fingerprint containing bits shared by V1FP and V2FP (CommV1V2FP), and two value fragment difference fingerprints (V1FP-V2FP and V2FP-V1FP, respectively). For keyed fingerprint descriptors such as MACCS[23] (in which each bit position is assigned to a specific structural fragment or pattern), a difference fingerprint of size $2n$ was calculated from value fingerprints of size n by merging the value fingerprints with uniquely set bit positions. For hashed fingerprints such as extended connectivity fingerprints (ECFPs),[24] which represent feature sets, a difference fingerprint was generated by combining unique features of the first value and unique features of the second value with inverted sign. These types of difference fingerprints accounted for direction-dependent transformations. To implement these five fingerprint designs, different fingerprint descriptors were used including ECFPs of bond diameter 2, 4, and 6 as well as MACCS structural keys (166 bit version). Fingerprint calculations were carried out using in-house Python scripts based upon the OEChem Toolkit.[21]

Support Vector Regression. SVR is a nonlinear prediction method and variant of SVM classification that maps data points into higher-dimensional chemical (descriptor) reference spaces with the aid of kernel functions. A linear regression is performed in the high dimensional space. More formally, the predicted value y of the input vector x is calculated as

$$y = K(w, x) + b$$

$K(w,x)$ is the kernel function that maps x into the high-dimensional space; w (the normal weight vector) and b (the bias) determine a hyperplane in this space to separate positive and negative training instances and are estimated by minimizing the so-called structural risk:

$$R_{\mathrm{SVR}}(C) = C\frac{1}{n}\sum_{n}^{i=1} L_{\varepsilon}(d_i, y_i) + \frac{1}{2}\|w\|^2$$

$$L_{\varepsilon}(d_i, y_i) = \begin{cases} |d - y| - \varepsilon & |d - y| \geq \varepsilon \\ 0 & \text{otherwise} \end{cases}$$

R_{SVR} represents the risk function and depends on a loss function L_ε and a regularization parameter depending on w. L_ε measures the difference between the predicted value y and observed value d and is only applied when the difference is larger than ε. C is a factor determining the trade-off between the loss function and the regularization term.

Kernel Functions. For SVR, six alternative kernels were designed and investigated. All kernels were based on the Tanimoto similarity function[25] but utilized different types of fingerprint representations. The design of these kernels was

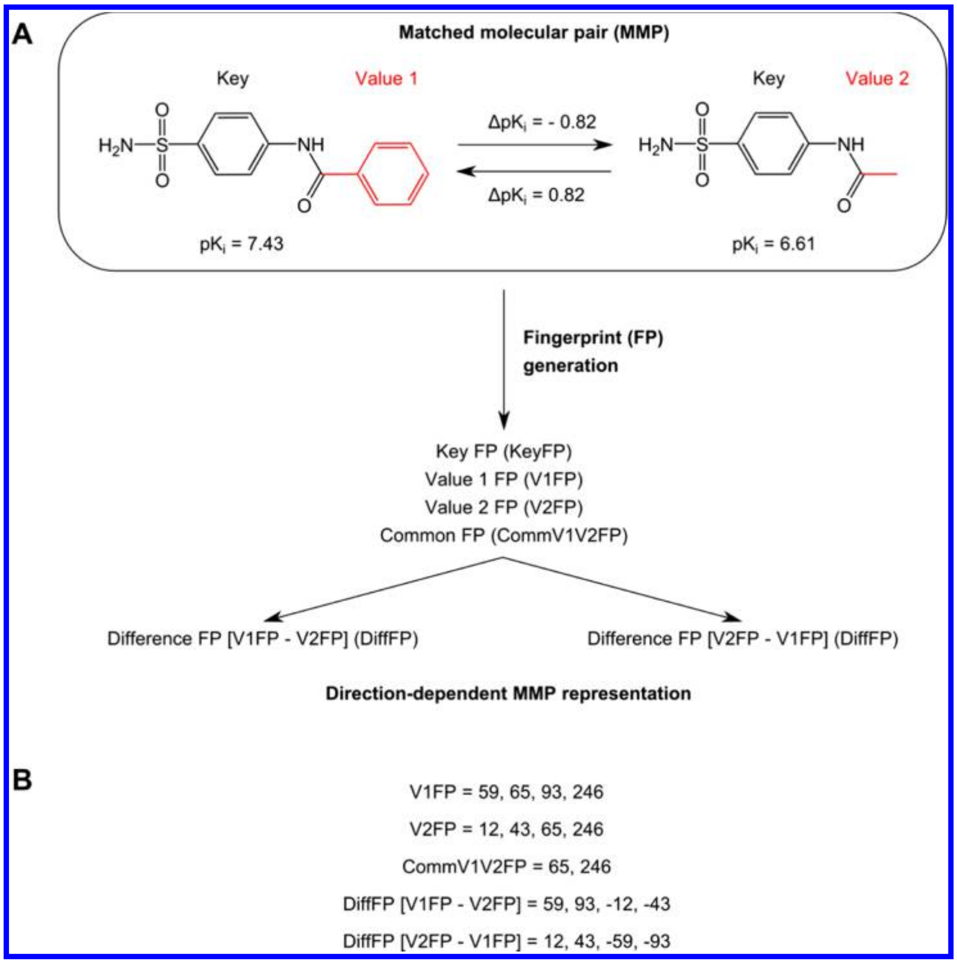

A
Matched molecular pair (MMP)
Key
Value 1
Key
Value 2
ΔpKi = - 0.82
ΔpKi = 0.82
pKi = 7.43
pKi = 6.61
Fingerprint (FP) generation
Key FP (KeyFP)
Value 1 FP (V1FP)
Value 2 FP (V2FP)
Common FP (CommV1V2FP)
Difference FP [V1FP - V2FP] (DiffFP)
Difference FP [V2FP - V1FP] (DiffFP)
Direction-dependent MMP representation
B
V1FP = 59, 65, 93, 246
V2FP = 12, 43, 65, 246
CommV1V2FP = 65, 246
DiffFP [V1FP - V2FP] = 59, 93, -12, -43
DiffFP [V2FP - V1FP] = 12, 43, -59, -93

Figure 1. Molecular representation. (A) At the top, an exemplary MMP is shown. The shared core structure (key fragment) is colored black, and the distinguishing value fragments (value 1 and value 2) are colored red. For potency prediction, the potency difference between MMP compounds is monitored in a direction-dependent manner. Accordingly, two direction-dependent MMPs are obtained encoding a potency-decreasing (value 1 → 2) and a reverse potency-increasing (value 2 → 1) transformation. For each MMP, a fingerprint of the key fragment (KeyFP) and a fingerprint of its value fragments (V1FP and V2FP, respectively) are calculated. From V1FP and V2FP, a common fingerprint (CommV1V2FP) is generated consisting of shared bit positions. In addition, two difference fingerprints (DiffFP) are calculated from V1FP and V2FP to yield a direction-dependent transformation representation. For direction-dependent MMPs, KeyFP and CommV1V2FP are conserved. (B) Schematic representation of feature set fingerprints of two value fragments and of CommV1V2FP and DiffFP derived from these feature sets. The derivation of CommV1V2FP and DiffFP is described in detail in the text.

inspired by MMP-based kernel functions successfully used for the prediction of activity cliffs.[18]

Transformation kernels only utilized value fragment-based fingerprints and represented a chemical transformation in three different ways, based upon

(i) only the difference fingerprint (DiffFP, Figure 1B)—this kernel was termed 1VD

(ii) both CommV1V2FP and DiffFP (2VCD)

(iii) the two value fragments fingerprints (2V12)

In addition, *MMP kernels* were constructed by adding the key fingerprint representation to i–iii producing kernels 2VKD, 3VKCD, and 3VK12, respectively.

These six kernels were implemented using the Tanimoto coefficient (Tc) formula and are defined by the following equations:

$$K_{1\mathrm{VD}}(i,j) = \mathrm{Tc}(\mathrm{DiffFP}_i, \mathrm{DiffFP}_j)$$

$$K_{2\mathrm{V12}}(i,j) = \mathrm{Tc}(\mathrm{V1FP}_i, \mathrm{V1FP}_j)\cdot\mathrm{Tc}(\mathrm{V2FP}_i, \mathrm{V2FP}_j)$$

$$K_{2\mathrm{VCD}}(i,j) = \mathrm{Tc}(\mathrm{DiffFP}_i, \mathrm{DiffFP}_j)\cdot\mathrm{Tc}(\mathrm{CommV1V2FP}_i, \mathrm{CommV1V2FP}_j)$$

$$K_{2\mathrm{VKD}}(i,j) = \mathrm{Tc}(\mathrm{KeyFP}_i, \mathrm{KeyFP}_j)\cdot K_{1\mathrm{VD}}(i,j)$$

$$K_{3\mathrm{VK12}}(i,j) = \mathrm{Tc}(\mathrm{KeyFP}_i, \mathrm{KeyFP}_j)\cdot K_{2\mathrm{V12}}(i,j)$$

$$K_{3\mathrm{VKCD}}(i,j) = \mathrm{Tc}(\mathrm{KeyFP}_i, \mathrm{KeyFP}_j)\cdot K_{2\mathrm{VCD}}(i,j)$$

where i and j are two direction-dependent MMPs and Tc(FP_i, FP_j) represents the Tc for comparison of the two fingerprints. MMP kernels were expected to project potency-annotated MMPs into features spaces in which linear modeling algorithms could be successfully applied. SVMlight[26] was used to perform all SVR calculations. Except for the kernel functions, default parameter settings were used.

Control Calculations. For comparison with SVR, two conceptually different approaches were used including MMP-based averaging analysis (MMPAV)[13,14] and RF predictions.[2,15] To predict the potency change of an MMP in the test set for

Table 2. SVR Predictions[a]

		A	B	C	D	E	F	G	H	I
1VD	MACCS	0.21	0.26	0.32	0.34	0.38	0.34	0.30	0.42	0.45
	ECFP2	0.32	0.35	0.43	0.43	0.43	0.41	0.36	0.51	0.53
	ECFP4	0.34	0.39	0.46	0.47	0.46	0.42	0.38	0.53	0.55
	ECFP6	0.34	0.40	0.46	0.47	0.46	0.42	0.39	0.53	0.55
2V12	MACCS	0.20	0.28	0.33	0.38	0.40	0.33	0.32	0.44	0.47
	ECFP2	0.27	0.35	0.43	0.41	0.43	0.38	0.36	0.50	0.51
	ECFP4	0.26	0.36	0.41	0.41	0.43	0.36	0.35	0.49	0.51
	ECFP6	0.25	0.35	0.40	0.40	0.41	0.36	0.34	0.48	0.50
2VCD	MACCS	0.19	0.28	0.33	0.36	0.39	0.33	0.31	0.43	0.46
	ECFP2	0.29	0.36	0.43	0.42	0.43	0.39	0.35	0.50	0.51
	ECFP4	0.30	0.38	0.44	0.44	0.45	0.40	0.36	0.51	0.53
	ECFP6	0.29	0.38	0.43	0.45	0.45	0.39	0.36	0.51	0.53
2VKD	MACCS	0.36	0.43	0.57	0.58	0.62	0.53	0.60	0.60	0.71
	ECFP2	0.49	0.58	0.73	0.67	0.74	0.64	0.71	0.73	0.79
	ECFP4	**0.50**	0.63	0.77	0.72	0.78	0.67	0.75	0.75	0.83
	ECFP6	0.49	**0.64**	**0.78**	**0.73**	**0.78**	**0.68**	**0.75**	**0.76**	**0.83**
3VK12	MACCS	0.28	0.38	0.52	0.52	0.58	0.47	0.54	0.56	0.66
	ECFP2	0.35	0.51	0.66	0.58	0.66	0.54	0.63	0.66	0.71
	ECFP4	0.33	0.53	0.66	0.58	0.66	0.53	0.64	0.65	0.72
	ECFP6	0.31	0.52	0.64	0.58	0.65	0.53	0.63	0.63	0.71
3VKCD	MACCS	0.28	0.39	0.52	0.51	0.58	0.47	0.54	0.55	0.65
	ECFP2	0.41	0.53	0.69	0.61	0.69	0.59	0.65	0.68	0.74
	ECFP4	0.41	0.57	0.71	0.65	0.71	0.60	0.69	0.69	0.77
	ECFP6	0.40	0.57	0.71	0.66	0.72	0.61	0.69	0.69	0.78

		J	K	L	M	N	O	P	Q
1VD	MACCS	0.38	0.37	0.48	0.60	0.24	0.40	0.43	0.18
	ECFP2	0.45	0.46	0.52	0.72	0.28	0.49	0.50	0.21
	ECFP4	0.47	0.49	0.55	0.74	0.29	0.52	0.53	0.22
	ECFP6	0.47	0.50	0.55	0.74	0.29	0.52	0.54	0.22
2V12	MACCS	0.40	0.40	0.50	0.64	0.24	0.42	0.44	0.15
	ECFP2	0.45	0.44	0.52	0.73	0.25	0.47	0.47	0.17
	ECFP4	0.44	0.44	0.53	0.72	0.24	0.46	0.48	0.17
	ECFP6	0.43	0.44	0.52	0.71	0.24	0.45	0.47	0.16
2VCD	MACCS	0.38	0.38	0.48	0.62	0.23	0.41	0.42	0.15
	ECFP2	0.44	0.44	0.52	0.72	0.26	0.47	0.48	0.19
	ECFP4	0.46	0.46	0.54	0.73	0.26	0.48	0.51	0.20
	ECFP6	0.46	0.47	0.54	0.73	0.26	0.49	0.51	0.19
2VKD	MACCS	0.63	0.63	0.70	0.79	0.44	0.67	0.61	0.37
	ECFP2	0.77	0.73	0.79	0.89	0.51	0.76	0.73	0.48
	ECFP4	0.80	0.77	0.83	0.91	0.55	0.81	0.77	0.53
	ECFP6	**0.81**	**0.79**	**0.83**	**0.91**	**0.57**	**0.81**	**0.77**	**0.54**
3VK12	MACCS	0.58	0.57	0.66	0.76	0.35	0.61	0.57	0.29
	ECFP2	0.69	0.64	0.72	0.85	0.38	0.67	0.64	0.36
	ECFP4	0.69	0.64	0.72	0.85	0.38	0.67	0.64	0.38
	ECFP6	0.67	0.63	0.71	0.84	0.39	0.65	0.63	0.38
3VKCD	MACCS	0.57	0.56	0.65	0.75	0.37	0.60	0.55	0.32
	ECFP2	0.72	0.67	0.75	0.86	0.44	0.70	0.67	0.42
	ECFP4	0.75	0.70	0.78	0.87	0.47	0.73	0.70	0.46
	ECFP6	0.75	0.71	0.78	0.87	0.48	0.73	0.71	0.47

[a] For each data set, average R^2 values after 10-fold cross-validation are reported for all fingerprint-kernel combinations. Overall best results are shown in bold.

MMPAV, the training set was searched for MMPs containing the same chemical transformation. If no such MMP was found, prediction was not possible. If qualifying training set MMPs were detected, the average potency difference of these MMPs was calculated to predict the potency change for the test MMP. RF modeling utilizes ensembles of decision trees for consensus predictions. RF calculations were performed using the R[27] package randomForest.[28] An MMP was represented as the difference of 51 2D numerical descriptors calculated with the Molecular Operating Environment (MOE)[29] and the potency value of the first compound of the MMP. The numerical descriptor set, which was not used for MMP-based SVM modeling, was previously designed for machine learning applications.[15] For RF calculations, the number of trees was set to 400; for all other parameters, default settings were used.

Table 3. SVR Results for the ECFP6-2VKD Combination (10-Fold Cross-Validation)[a]

	R^2	SD(R^2)	MAE	SD(MAE)	RMSE	SD(RMSE)	r	SD(r)
A	0.49	0.06	0.48	0.04	0.84	0.10	0.73	0.04
B	0.64	0.06	0.30	0.02	0.48	0.06	0.81	0.04
C	0.78	0.01	0.23	0.01	0.37	0.01	0.89	0.01
D	0.73	0.02	0.34	0.01	0.50	0.02	0.86	0.01
E	0.78	0.02	0.24	0.01	0.37	0.03	0.89	0.01
F	0.68	0.03	0.31	0.01	0.47	0.03	0.84	0.02
G	0.75	0.03	0.29	0.02	0.46	0.03	0.87	0.02
H	0.76	0.02	0.30	0.02	0.47	0.03	0.88	0.01
I	0.83	0.02	0.27	0.01	0.41	0.02	0.91	0.01
J	0.81	0.02	0.24	0.01	0.38	0.02	0.91	0.01
K	0.79	0.03	0.33	0.02	0.50	0.04	0.89	0.02
L	0.83	0.02	0.29	0.01	0.46	0.02	0.92	0.01
M	0.91	0.01	0.18	0.00	0.28	0.01	0.96	0.00
N	0.57	0.04	0.47	0.03	0.73	0.05	0.77	0.03
O	0.81	0.02	0.22	0.01	0.35	0.02	0.91	0.01
P	0.77	0.04	0.27	0.02	0.40	0.02	0.89	0.02
Q	0.54	0.06	0.35	0.03	0.56	0.06	0.75	0.04

[a] For each data set, results of SVR predictions for the preferred ECFP6−2VKD combination are reported using different performance measures including the average and standard deviation (SD) of the coefficient of determination (R^2), mean absolute error (MAE), root-mean-square error (RMSE), and correlation coefficient (r) over 10 independent trials.

The RF protocol followed a previous report of predictions of MMP-encoded property changes.[15]

In order to evaluate if prediction accuracy resulted from contributions of newly designed kernels or was essentially determined by SVR, the performance of SVR calculations using the overall preferred kernel/fingerprint combination was compared to kernel ridge regression (KRR),[30] which represents an alternative kernel-based regression method. Therefore, the MMP kernels were implemented in R using the package kernlab,[31] and KRR calculations were carried out using the R package CVST.[32] Except for the kernel function, default parameter settings were used.

Learning and Scoring. Training calculations for SVR, RF, and MMPAV were based on 10-fold cross-validation. Compound pairs forming direction-dependent MMPs were randomly divided into 10 nonoverlapping groups. For each compound pair, the two direction-dependent MMPs were assigned to the same group. Regression was performed 10 times. Each time a different group was chosen as the test set, and the remaining nine groups were combined as the training set.

As further control calculations, SVR was also performed using 4-fold cross-validation on the basis of four data set subsets (instead of 10) and, in addition, without cross-validation on larger test sets (i.e., randomly selecting half of each data set as the training and the other half as the test set). These calculations were carried out in order to evaluate the influence of training set composition and size on prediction accuracy. For comparing KRR and SVR, no cross-validation was applied due to the high computational expense of KRR calculations.

For each prediction, the coefficient of determination (R^2), mean absolute error (MAE), root-mean-square error (RMSE), and Pearson's correlation coefficient (r) were calculated comparing the predicted and the observed potency difference values for the test set. Following cross-validation, the average and standard deviation (SD) of the scores for each independent trial were calculated and used as the final prediction result. The different performance measures applied herein are defined by the following equations:

$$R^2 = 1 - \frac{\sum (\text{obs}_i - \text{pred}_i)^2}{\sum (\text{obs}_i - \overline{\text{obs}})^2}$$

$$\text{MAE} = \frac{\sum |\text{obs}_i - \text{pred}_i|}{N}$$

$$\text{RMSE} = \sqrt{\frac{\sum (\text{obs}_i - \text{pred}_i)^2}{N}}$$

$$r = \frac{\sum (\text{obs}_i - \overline{\text{obs}})\cdot(\text{pred}_i - \overline{\text{pred}})}{\sqrt{\sum (\text{obs}_i - \overline{\text{obs}})^2 \cdot \sum (\text{pred}_i - \overline{\text{pred}})^2}}$$

■ RESULTS AND DISCUSSION

Kernel Characteristics. Newly designed kernel functions accounted for transformation and MMP information in

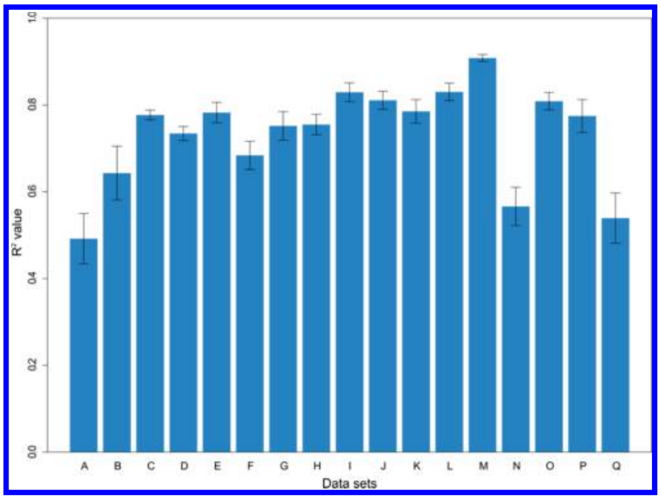


Figure 2. Best SVR predictions. For each data set, the R^2 value for the ECFP6−2VKD fingerprint−kernel combination is reported. R^2 standard deviations for 10 independent trials are given at the top of each bar.

Table 4. SVR Results for the ECFP6-2VKD Combination (4-Fold Cross-Validation)[a]

	R^2	SD(R^2)	MAE	SD(MAE)	RMSE	SD(RMSE)	r	SD(r)
A	0.45	0.03	0.50	0.03	0.88	0.09	0.71	0.02
B	0.61	0.02	0.32	0.01	0.51	0.02	0.79	0.01
C	0.75	0.01	0.24	0.00	0.39	0.01	0.88	0.01
D	0.72	0.01	0.35	0.01	0.52	0.01	0.86	0.01
E	0.76	0.01	0.25	0.01	0.38	0.01	0.88	0.01
F	0.66	0.02	0.32	0.01	0.48	0.01	0.83	0.01
G	0.74	0.02	0.30	0.00	0.47	0.01	0.87	0.01
H	0.74	0.01	0.31	0.01	0.49	0.01	0.87	0.01
I	0.82	0.00	0.28	0.01	0.43	0.01	0.91	0.00
J	0.79	0.01	0.26	0.01	0.40	0.01	0.90	0.00
K	0.78	0.02	0.34	0.00	0.51	0.02	0.89	0.01
L	0.81	0.01	0.31	0.00	0.48	0.01	0.91	0.01
M	0.90	0.01	0.19	0.00	0.29	0.01	0.95	0.00
N	0.54	0.02	0.49	0.01	0.75	0.03	0.75	0.01
O	0.78	0.01	0.23	0.00	0.37	0.01	0.90	0.00
P	0.75	0.01	0.29	0.01	0.43	0.01	0.88	0.01
Q	0.50	0.02	0.36	0.02	0.58	0.02	0.72	0.02

[a] For each data set, results of SVR predictions for the preferred ECFP6−2VKD combination are reported using different performance measures including the average and standard deviation (SD) of the coefficient of determination (R^2), mean absolute error (MAE), root-mean-square error (RMSE), and correlation coefficient (r) over four independent trials.

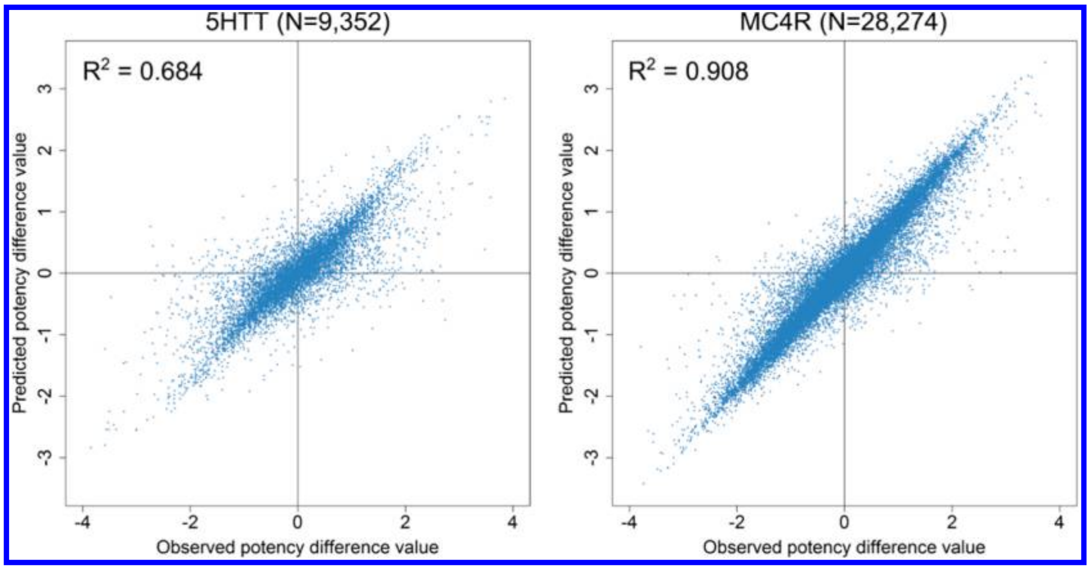


Figure 3. Comparison of predicted and observed potency difference values. For two exemplary data sets, 5HTT and MC4R, predicted and observed potency value differences are compared in a scatterplot. R^2 values and the number of MMPs (N) plotted are given. Each data point represents a direction-dependent MMP.

different ways. For example, CommV1V2FP and DiffFPs generated for value fragments and the kernels built using these representations took combined transformation information and/or structural differences between exchanged substructures into account. By contrast, the incorporation of KeyFPs added core structure information to transformation kernels and hence represented the structural context in which transformations occurred. The design of MMP pair product kernels is based upon pre-existing pairwise Tanimoto kernel products.

Systematic SVR Predictions. For each data set, 24 SVR predictions were carried out resulting from combinations of four fingerprint descriptors (ECFP2, ECFP4, ECFP6, and MACCS) and six kernel functions (1VD, 2V12, 2VCD, 2VKD, 3VK12, and 3VKCD), as described in detail in the Methods section. The R^2 results of 10-fold cross-validated calculations are reported in Table 2. R^2 values significantly varied for different combinations and data sets and ranged from 0.15 to 0.91, hence reflecting significant differences in prediction accuracy. However, regardless of the magnitude of R^2 values, most trials produced stable results with low R^2 standard deviations of, on average, only 0.037.

Kernel and Descriptor Performance. For all data sets, MMP kernels (taking core structure and transformation information into account) were found to perform better than transformation kernels. The average R^2 value for all calculations with MMP kernels was 0.63 compared to 0.42 for transformation kernels. Thus, structural context information was of critical importance for accurate potency difference value predictions associated with specific chemical transformations. Furthermore, the best representation of a transformation was the value difference fingerprint (1VD and the combined 2VKD

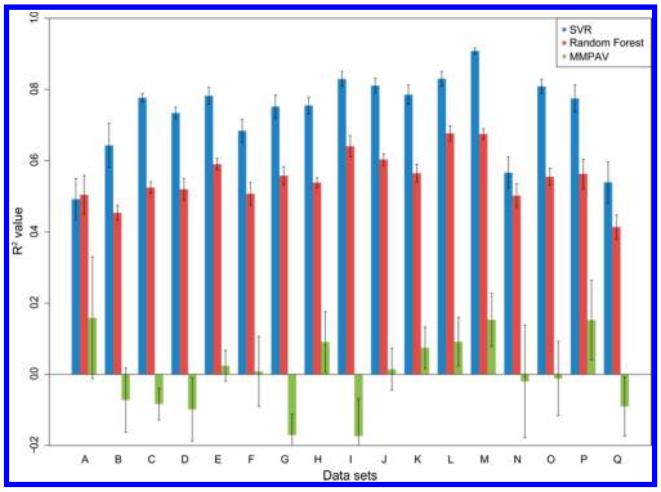


Figure 4. Control calculations. For each data set, the performance of SVR using the ECFP6−2VKD combination (blue) is compared to RF (red) and MMPAV (green) calculations. R^2 values and standard deviations over 10 independent trials are reported.

version), rather than the common plus difference fingerprints (2VCD and 3VKCD) or the two individual value fingerprints (2V12 and 3VK12). The average R^2 for 1VD (0.43) was slightly larger than for 2VCD (0.41) and 2V12 (0.40). Similarly, the average R^2 for 2VKD (0.68) was larger than for 3VKCD (0.62) and 3VK12 (0.58). Hence, kernel 2VKD, which combined the value difference fingerprint and the key fingerprint, performed overall best.

ECFP fingerprints consistently yielded higher performance than MACCS structural keys. Furthermore, increasing the diameter of the ECFPs (and hence their topological resolution) improved SVR prediction performance in most cases, albeit with generally low differences between ECFP4 and ECFP6.

Preferred Combination. The combination of the 2VKD kernel and the high-resolution ECFP6 gave the maximum R^2 value among all 24 kernel-fingerprint combinations for 16 of 17 compound data sets (Table 2). In Table 3, prediction results are reported for 10-fold cross-validation for the preferred 2VKD−ECFP6 combination. R^2 values varied between 0.49 (CA2) and 0.91 (MC4R), as also shown in Figure 2. For 11 of 17 compound data sets, R^2 values of at least 0.75 were obtained (Tables 2 and 3), reflecting generally high prediction accuracy. Again, R^2 standard deviations for these calculations were very low (Table 3), ranging from 0.01 (MC4R) to 0.06 (5-HT1A). Alternative performance measures were applied. MAE and RMSE results showed trends similar to those observed for R^2. MAE values varied between 0.18 (MC4R) and 0.48 (CA2) and RMSE values between 0.28 (MC4R) and 0.84 (CA2). The correlation between observed and predicted values was generally high, even for predictions yielding intermediate R^2 values. Only two compound data sets had correlation coefficient (r) values below 0.8 and five data sets had values above 0.9. In Table 4, prediction results are reported for 4-fold cross-validation for 2VKD−ECFP6 combination. R^2 values were very similar to those obtained with 10-fold cross-validation (Table 3) showing that test set size and composition had no major influence on prediction accuracy.

Figure 3 shows scatterplots comparing observed and predicted potency difference values for two exemplary data sets with moderate (5HTT) and high (MC4R) prediction accuracy. The observed potency differences reported in these plots are also representative for the compound data sets under study. Across all data sets, only ∼4.5% of the direction-dependent MMPs encoded potency differences of 2 orders of magnitude or more. The comparison in Figure 3 revealed that predicted potency differences covered the entire range of observed differences. Moreover, the pronounced diagonal patterns resulted from the presence of many highly accurate predictions.

Control Calculations. To put SVR performance into perspective, control calculations were carried out using MMPAV and RF. Especially RF calculations were relevant for comparison with SVR, given a previous report that utilized RF analysis to predict MMP-associated changes in property values.[15] Figure 4 reports the results of control calculations compared to SVR using the preferred 2VKD−ECFP6 combination. MMPAV performed poorly and even produced negative R^2 for eight compound sets. Details are provided in Table 5. Between 35% (CA9) and 70% (MC4R) of the test

Table 5. MMP-Based Averaging Analysis[a]

	R^2	SD(R^2)	MAE	SD(MAE)	RMSE	SD(RMSE)	R	SD(R)
A	0.16	0.17	0.69	0.03	0.98	0.04	0.47	0.14
B	−0.07	0.09	0.60	0.03	0.81	0.06	0.26	0.09
C	−0.08	0.04	0.60	0.03	0.81	0.03	0.29	0.04
D	−0.10	0.09	0.67	0.02	0.90	0.02	0.29	0.04
E	0.02	0.04	0.53	0.02	0.73	0.03	0.38	0.03
F	0.01	0.10	0.57	0.04	0.76	0.05	0.35	0.06
G	−0.17	0.06	0.69	0.02	0.93	0.04	0.24	0.04
H	0.09	0.09	0.64	0.03	0.86	0.04	0.44	0.05
I	−0.17	0.11	0.67	0.02	0.90	0.03	0.26	0.06
J	0.01	0.06	0.61	0.02	0.83	0.04	0.38	0.04
K	0.07	0.06	0.70	0.03	0.93	0.05	0.42	0.04
L	0.09	0.07	0.69	0.03	0.95	0.04	0.45	0.04
M	0.15	0.07	0.51	0.02	0.70	0.03	0.52	0.04
N	−0.02	0.16	0.75	0.08	1.06	0.10	0.33	0.11
O	−0.01	0.10	0.51	0.03	0.71	0.04	0.38	0.07
P	0.15	0.11	0.48	0.03	0.65	0.04	0.46	0.07
Q	−0.09	0.08	0.60	0.06	0.85	0.09	0.28	0.04

[a] For each data set, results of MMPAV calculations are reported using different performance measures according to Table 3.

Table 6. Random Forest Control Calculations[a]

	R^2	SD(R^2)	MAE	SD(MAE)	RMSE	SD(RMSE)	r	SD(r)
A	0.50	0.05	0.57	0.03	0.83	0.08	0.71	0.04
B	0.45	0.02	0.43	0.02	0.60	0.04	0.68	0.02
C	0.52	0.02	0.39	0.01	0.53	0.01	0.73	0.01
D	0.52	0.03	0.52	0.02	0.67	0.02	0.72	0.02
E	0.59	0.02	0.37	0.01	0.51	0.02	0.77	0.01
F	0.51	0.03	0.44	0.02	0.58	0.02	0.72	0.02
G	0.56	0.03	0.45	0.01	0.61	0.02	0.75	0.02
H	0.54	0.01	0.49	0.02	0.65	0.02	0.74	0.01
I	0.64	0.03	0.44	0.02	0.60	0.02	0.81	0.02
J	0.60	0.02	0.40	0.01	0.55	0.02	0.78	0.01
K	0.56	0.03	0.54	0.02	0.71	0.03	0.76	0.02
L	0.68	0.02	0.47	0.01	0.63	0.01	0.83	0.01
M	0.68	0.01	0.39	0.01	0.52	0.01	0.83	0.01
N	0.50	0.03	0.57	0.02	0.78	0.03	0.71	0.02
O	0.55	0.02	0.39	0.01	0.53	0.02	0.75	0.02
P	0.56	0.04	0.42	0.01	0.56	0.02	0.75	0.03
Q	0.41	0.03	0.44	0.03	0.63	0.05	0.64	0.03

[a] For each data set, results of RF calculations are reported using different performance measures according to Table 3.

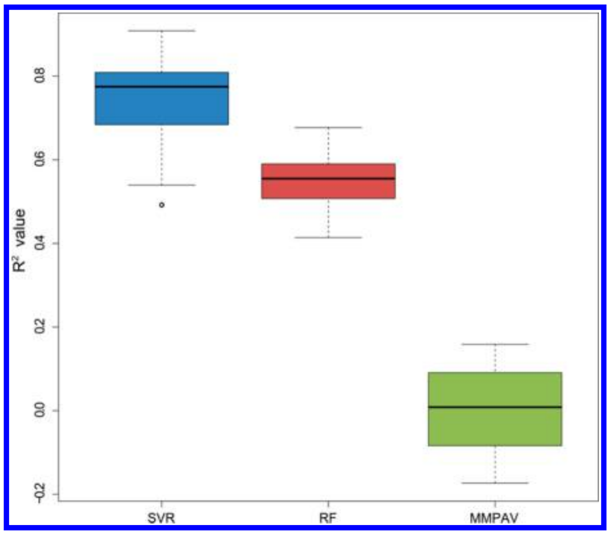


Figure 5. R^2 comparison. The R^2 value distributions resulting from the preferred SVR combination (blue), RF (red), and MMPAV (green) calculations are compared in a boxplot representation.

MMPs could not be predicted using MMPAV because a qualifying transformation was not available in the learning set. It should be noted that negative R^2 indicated that better predictions would be obtained by using the mean of the entire potency change distribution of a data set (essentially corresponding to random predictions). This very low performance was not unexpected for a simple averaging method. However, the results clearly indicated nonlinearity of many MMP-encoded SARs.

RF calculations yielded much better prediction performance than MMPAV. For most data sets, R^2 values between 0.5 and 0.6 were observed for RF predictions, with a maximum value of 0.68 (MC4R and ADORA3). Details are provided in Table 6. However, as reported in Figure 4, RF predictions did not reach the prediction accuracy of SVR for 16 of 17 compound data sets. Figure 5 compares R^2 values for the different calculations in boxplots and shows that the interquartile range of RF was lower than SVR and that the median R^2 value of SVR was ~0.2 units higher.

Table 7. Comparison of SVR and KRR for the ECFP6-2VKD Combination[a]

	R^2		MAE		RMSE		r	
	SVR	KRR	SVR	KRR	SVR	KRR	SVR	KRR
A	0.35	0.52	0.56	0.51	0.96	0.82	0.62	0.73
B	0.53	0.59	0.36	0.33	0.56	0.52	0.74	0.77
C	0.68	0.72	0.27	0.24	0.43	0.40	0.84	0.86
D	0.64	0.69	0.41	0.38	0.59	0.55	0.82	0.84
E	0.70	0.74	0.29	0.26	0.44	0.41	0.85	0.87
F	0.60	0.63	0.36	0.34	0.52	0.50	0.79	0.80
G	0.68	0.72	0.35	0.32	0.53	0.49	0.83	0.86
H	0.67	0.72	0.37	0.33	0.56	0.52	0.83	0.86
I	0.77	0.80	0.32	0.29	0.48	0.44	0.88	0.90
J	0.73		0.30		0.45		0.87	
K	0.70	0.75	0.41	0.36	0.58	0.54	0.85	0.87
L	0.75		0.36		0.55		0.87	
M	0.86		0.22		0.33		0.93	
N	0.48	0.57	0.54	0.49	0.80	0.73	0.71	0.76
O	0.71	0.76	0.27	0.24	0.43	0.39	0.86	0.88
P	0.69	0.74	0.34	0.31	0.47	0.43	0.84	0.87
Q	0.44	0.49	0.40	0.38	0.61	0.59	0.67	0.70

[a] For each data set, results of SVR and KRR predictions for the preferred ECFP6–2VKD combination are reported using different performance measures including the coefficient of determination (R^2), mean absolute error (MAE), root-mean-square error (RMSE), and correlation coefficient (r). For data sets J, L, and M, KRR calculations could not be completed.

Kernel ridge regression was compared to SVR using the preferred 2VKD–ECFP6 combination without cross-validation (i.e., on larger test sets). KRR is computationally much more demanding than SVR both in terms of memory requirements and in CPU time (KRR calculations could not be completed for three large data sets). Results are reported in Table 7. Compared to SVR, KRR calculations yielded a small increase in R^2 values for most data sets. These findings indicated that the newly designed kernels were largely responsible for the observed prediction accuracy, rather than SVR optimization details.

CONCLUSIONS

In this work, we have explored support vector regression using newly designed kernel functions for the prediction of numerical potency differences between compounds forming MMPs. Application of the MMP formalism for potency prediction further expands the applicability domain of QSAR-type approaches. This is the case because (*i*) many different structural relationships are captured at the level of compound pairs and (*ii*) MMPs encode well-defined chemical transformations in different structural environments. For potency difference prediction, the MMP approach was further refined by introducing direction-dependent MMPs. By combining MMP-based transformation analysis and machine learning approaches such as SVR, nonlinear SARs can be captured in structurally heterogeneous data sets. In our calculations, overall high SVR prediction accuracy was achieved for a preferred combination of a kernel taking transformation and core structure information into account and a high-resolution topological fingerprint descriptor. Transformation information was best captured by a fingerprint representation accounting for structural differences between the exchanged substructures. Given that potency difference values were predicted using SVR with reasonable to high accuracy for structurally analogous compounds from many different data sets, the methodology introduced herein should merit further consideration for compound potency predictions to complement and potentially extend existing QSAR approaches.

AUTHOR INFORMATION

Corresponding Author

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

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We are grateful to OpenEye Scientific Software, Inc., for the free academic license of the OpenEye Toolkits.

REFERENCES

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

(2) Breiman, L. Random Forests. *Machine Learning* **2001**, *45*, 5−32.

(3) Drucker, H.; Burges, C. Support Vector Regression Machines. *Adv. Neural Inform. Process. Systems* **1997**, *9*, 155−161.

(4) Yuan, Y.; Zhang, R.; Hu, R.; Ruan, X. Prediction of CCR5 Receptor Binding Affinity of Substituted 1-(3,3-diphenylpropyl)-piperidinyl Amides and Ureas Based on the Heuristic Method, Support Vector Machine and Projection Pursuit Regression. *Eur. J. Med. Chem.* **2009**, *44*, 25−34.

(5) Xue, C. X.; Zhang, R. S.; Liu, H. X.; Yao, X. J.; Liu, M. C.; Hu, Z. D.; Fan, B. T. QSAR Models for the Prediction of Binding Affinities to Human Serum Albumin Using the Heuristic Method and a Support Vector Machine. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1693−1700.

(6) Sun, M.; Chen, J.; Wei, H.; Yin, S.; Yang, Y.; Ji, M. Quantitative Structure-activity Relationship and Classification Analysis of Diaryl Ureas Against Vascular Endothelial Growth Factor Receptor-2 Kinase Using Linear and Non-linear Models. *Chem. Biol. Drug Des.* **2009**, *73*, 644−654.

(7) Lind, P.; Maltseva, T. Support Vector Machines for the Estimation of Aqueous Solubility. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1855−1859.

(8) Gombar, V. K.; Hall, S. D. Quantitative Structure-activity Relationship Models of Clinical Pharmacokinetics: Clearance and Volume of Distribution. *J. Chem. Inf. Model.* **2013**, *53*, 948−957.

(9) Fatemi, M. H.; Gharaghani, S. A Novel QSAR Model for Prediction of Apoptosis-inducing Activity of 4-aryl-4-H-chromenes Based on Support Vector Machine. *Bioorg. Med. Chem.* **2007**, *15*, 7746−7754.

(10) Leong, M. K. A Novel Approach Using Pharmacophore Ensemble/Support Vector Machine (PhE/SVM) for Prediction of hERG Liability. *Chem. Res. Toxicol.* **2007**, *20*, 217−226.

(11) Song, M.; Clark, M. Development and Evaluation of an in Silico Model for hERG Binding. *J. Chem. Inf. Model.* **2005**, *46*, 392−400.

(12) Kenny, P. W.; Sadowski, J. Structure Modification in Chemical Databases. In *Chemoinformatics in Drug Discovery*; Oprea, T. I., Ed.; Wiley-VCH: Weinheim, Germany, 2005; pp 271−285.

(13) Sheridan, R. P.; Hunt, P.; Culberson, J. C. Molecular Transformations as a Way of Finding and Exploiting Consistent Local QSAR. *J. Chem. Inf. Model.* **2006**, *46*, 180−192.

(14) de la Vega de León, A.; Bajorath, J. Compound Optimization Through Data Set-dependent Chemical Transformations. *J. Chem. Inf. Model.* **2013**, *53*, 1263−1271.

(15) Beck, J. M.; Springer, C. Quantitative Structure-activity Relationship Models of Chemical Transformations from Matched Pairs Analyses. *J. Chem. Inf. Model.* **2014**, *54*, 1226−1234.

(16) Stumpfe, D.; Hu, Y.; Dimova, D.; Bajorath, J. Recent Progress in Understanding Activity Cliffs and Their Utility in Medicinal Chemistry. *J. Med. Chem.* **2014**, *57*, 18−28.

(17) Cortes, C.; Vapnik, V. Support-vector Networks. *Machine Learning* **1995**, *20*, 273−297.

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

(19) 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*, D1100−1107.

(20) Hussain, J.; Rea, C. Computationally Efficient Algorithm to Identify Matched Molecular Pairs (MMPs) in Large Data Sets. *J. Chem. Inf. Model.* **2010**, *50*, 339−348.

(21) *OEChem*, v. Feb2014; OpenEye Scientific Software Inc: Santa Fe, NM, 2014.

(22) Hu, X.; Hu, Y.; Vogt, M.; Stumpfe, D.; Bajorath, J. MMP-Cliffs: Systematic Identification of Activity Cliffs on the Basis of Matched Molecular Pairs. *J. Chem. Inf. Model.* **2012**, *52*, 1138−1145.

(23) *MACCS Structural Keys*; Symyx Software: San Ramon, CA, 2005.

(24) Rogers, D.; Hahn, M. Extended-connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.

(25) Willett, P.; Barnard, J.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.

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

(27) *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2008.

(28) Liaw, A.; Wiener, M. Classification and Regression by randomForest. *R News* **2002**, *2*, 18−22.

(29) *Molecular Operating Environment (MOE)*; Chemical Computing Group Inc.: Montreal, Canada, 2011.

(30) Christianini, N.; Shawe-Taylor, J. *An Introduction to Support Vector Machines and other Kernel-based Learning Methods*; Cambridge University Press: Cambridge, UK, 2000.

(31) Karatzoglou, A.; Smola, A.; Hornik, K.; Zeileis, A. kernlab - An S4 Package for Kernel Methods in R. *J. Stat. Softw.* **2004**, *11*, 1−20.

(32) CVST R package. http://cran.r-project.org/web/packages/CVST/index.html.