JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

Beyond the Scope of Free-Wilson Analysis: Building Interpretable QSAR Models with Machine Learning Algorithms

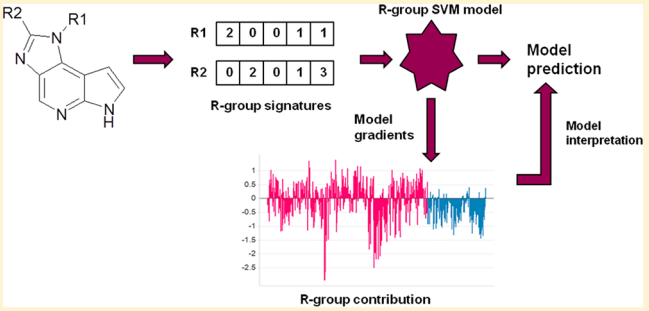
Hongming Chen,*[,†] Lars Carlsson,[‡] Mats Eriksson,[†] Peter Varkonyi,[†] Ulf Norinder,[¥,&] and Ingemar Nilsson*[,£]

[†]Chemistry Innovation Center, Discovery Sciences, [‡]Computational Toxicology, Global Safety Assessment and [£]CVGI Innovative Medicines, AstraZeneca R&D Mölndal, Sweden

[¥]CNSP Innovative Medicines, AstraZeneca R&D Södertälje, Sweden

Supporting Information

ABSTRACT: A novel methodology was developed to build Free-Wilson like local QSAR models by combining R-group signatures and the SVM algorithm. Unlike Free-Wilson analysis this method is able to make predictions for compounds with R-groups not present in a training set. Eleven public data sets were chosen as test cases for comparing the performance of our new method with several other traditional modeling strategies, including Free-Wilson analysis. Our results show that the R-group signature SVM models achieve better prediction accuracy compared with Free-Wilson analysis in general. Moreover, the predictions of R-group signature models are also comparable to the models using ECFP6 fingerprints and signatures for the whole compound. Most importantly, R-group contributions to the SVM model can be obtained by calculating the gradient for R-group signatures. For most of the studied data sets, a significant correlation with that of a corresponding Free-Wilson analysis is shown. These results suggest that the R-group contribution can be used to interpret bioactivity data and highlight that the R-group signature based SVM modeling method is as interpretable as Free-Wilson analysis. Hence the signature SVM model can be a useful modeling tool for any drug discovery project.



INTRODUCTION

The pharmaceutical industry constantly strives to decrease the time from lead generation through lead optimization (LO) to clinical development as well as finding ways to improve the quality of the compounds progressed.[1−3] The drug discovery project team needs to constantly reflect on incoming data and judge whether these will change the order of ranking of current best compounds and reprioritize project activities accordingly. The concomitant design and synthesis of new compounds is a complex, time-consuming task, in which the medicinal chemist seeks to balance potency, off-target interactions, pharmacokinetic properties, and toxicity. A critical step in project progression is to select the right compounds to synthesize from an almost infinite number of virtual compounds. Thus, computational models that predict compounds with the right properties are highly desirable. Quantitative structure−activity relationship (QSAR) or quantitative structure−property relationship (QSPR) approaches approximate the function between molecular properties and the corresponding biological end points of interest using multivariate statistics.[4] The molecular properties can be 1D (e.g., molecular weight, logP, property count, or structural descriptors), 2D (e.g., structural keys or hashed fingerprints), or 3D descriptors (CoMFA,[5] pharmacophore fingerprints). Linear (e.g., Partial Least Square and Multi Linear Regression) and nonlinear methods (e.g., Random Forest, Neural Network, and Support Vector Machines) have been extensively used to develop QSAR, and numerous models have been built to explain and predict biological activity.[6−11] The Free-Wilson approach was one of the first mathematical techniques developed for the QSAR for a series of chemical analogues.[12−14] It does not require any substituent parameters to be defined, and only the biological activity is needed. This is in contrast to Hansch analysis,[15,16] where physicochemical properties are correlated with biological activity values. The basic idea in the Free-Wilson approach is that the biological activity of a molecule can be described as the sum of the activity contributions of its specific substructures (parent core and the corresponding substituents (R-groups)), and the major advantage of this method is its interpretability. Although the Free-Wilson methodology has provided fruitful transparent models[17−24] for a range of experimental observations, it is hampered by the limitations in prediction scope. The training set needs to be properly designed to be able to predict a modeled property in an enumerated library for all R-group combinations, and naturally it does not provide explorative prediction for novel R-groups which are not present in the training set. Hence, we have explored bitmap fingerprint based

Received: March 4, 2013
Published: May 27, 2013

Downloaded via UNIVERSITA DI ROMA LA SAPIENZA on February 21, 2021 at 20:18:41 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

 © 2013 American Chemical Society dx.doi.org/10.1021/ci4001376 | J. Chem. Inf. Model. 2013, 53, 1324−1336

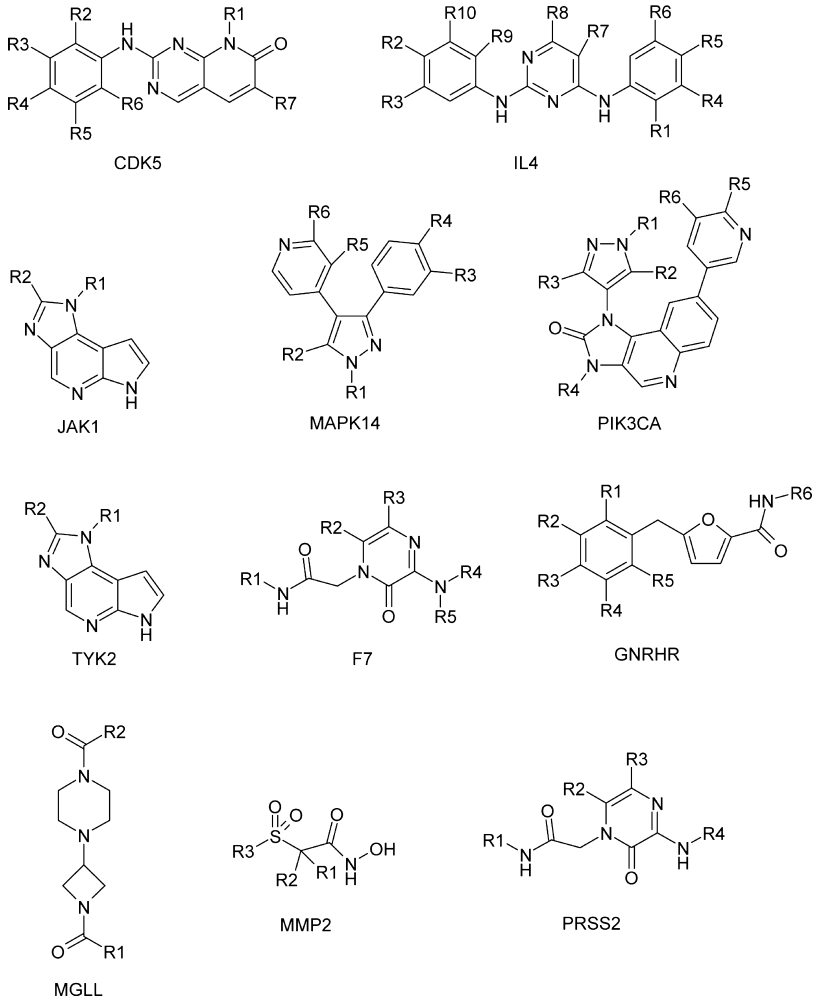


Figure 1. The scaffolds and R-groups for the data sets used in the study.

methods that can extend the scope of Free-Wilson models without forfeiting its interpretability. Molecular fingerprints are representations of chemical structures originally designed to assist in chemical database substructure searching, similarity searching, nearest neighbor analysis, clustering and classifications. DayLight,[25] Unity,[26] MDL[27] fingerprints, and the more recently developed Pipeline Pilot extended-connectivity fingerprint (ECFP)[28] are examples of this type of descriptor. Besides these commercially available fingerprint methodologies, there is a freely available atom signature descriptor (signatures) developed by Faulon et al.[29] Signatures describe, similar to ECFPs, circular fragments around each atom in a molecule. The circular fragments are represented as trees described by strings similar to the SMILES notation, but they are also canonical which allows for direct comparison between fragments within different molecules regardless of how the fragments are represented.

In the present study, we describe a new methodology to combine the support vector machine (SVM) algorithm and signature descriptors generated for R-groups around a common scaffold. This methodology inherits the key feature of Free-Wilson analysis, which is decomposition of bioactivity into R-group contributions and at the same time overcomes the prediction scope problem of the Free-Wilson method. Nonlinear SVM models are often regarded as noninterpretable black box models.[30] However, based on our previous work on interpreting machine learning models,[31] we have derived the concept of signature-based model gradient and utilize this as a measurement of signature contribution in the model. Here we report the QSAR model performance on eleven different data sets. The chosen data sets were retrieved from the GOSTAR databases[32] and were selected to represent different target classes, with consistent biological data from the same source and a sufficient number of data points to allow for solid statistics. At the same time, we also compared the predictive power of models using the Free-Wilson methodology, various SVM models using ECFP fingerprints, signatures generated from either full compounds or R-groups and traditional descriptors which represent bulk molecular physicochemical properties. Our study shows that fingerprint/signature based SVM models in general have higher prediction accuracy than Free-Wilson models. Comparing with SVM models generated from molecular signatures (of the full compound), a benefit of Free-Wilson like SVM models employing R-group signatures is that the relative contribution from individual R-groups to the observed activity can be derived and therefore provide the similar interpretability as the Free-Wilson method.

METHODS

Data Set. In total, eleven experimental data sets were used in the current study, taken from published chemical patents and extracted from the GOSTAR database.[32] The scaffolds of these

dx.doi.org/10.1021/ci4001376 | J. Chem. Inf. Model. 2013, 53, 1324−1336

Table 1. Data Sets Used in the Study

data set	compound description	data type	nr. training set	nr. test set	data source
CDK5	CDK5 inhibitor	IC50	184	46	US 20040224958 A1
IL4	IL-4 inhibitor	IC50	532	133	WO 2006/133426 A2
JAK1	JAK1 inhibitor	K_i	736	185	WO 2011/086053 A1
MAPK14	P38 alpha inhibitor	IC50	488	122	EP 1500657 A1
PIK3CA	PI3K alpha inhibitor	IC50	243	61	WO 2010/139731 A1
TYK2	TYK2 inhibitor	K_i	736	184	WO 2011/086053 A1
F7	factor VIIa inhibitor	IC50	292	73	US20050043313
GNRHR	gonadotropin-releasing hormone receptor antagonist	IC50	159	39	WO20020358
MGLL	monoacylglycerol lipase inhibitor	IC50	982	246	WO2010124082
MMP2	matrix metalloprotease 2 inhibitor	IC50	439	110	WO2005042521
PRSS2	trypsin II inhibitor	IC50	271	68	US7119094

eleven compound libraries are shown in Figure 1, and detailed information for each data set is listed in Table 1. Each data set was randomly split into training and test sets with a ratio of 4:1 by a Perl script[33] for model validation purposes.

Modeling Methods. Two different molecular fingerprint methods were used in the current study. One fingerprint method is the molecular signature developed by Faulon et al.[29] The signature descriptors were calculated using an internal AstraZeneca implementation based on the publically available Openbabel toolkit.[34] The generated signature for each structure is a sparse vector, in which each bit is an integer representing the occurrence of a specific signature in a compound or R-group. Atom signatures were computed for each molecule in a data set of heights from zero to three. Thus for each atom in a molecule four atom signatures were generated. Both full molecular signatures and R-groups signatures were generated to compare their performance. The R-group signatures serve as a direct comparison with the Free-Wilson method, since both methods are characterizing the R-groups for a focused library. The pipeline pilot ECFP6 fingerprint[35] with 1024 bits was used as another fingerprint method to build QSAR models. For ECFP6, only full compound fingerprints were generated for model building. Besides the fingerprint based descriptors, a set of bulk property based descriptors which comprise 196 2D/3D descriptors (referred to as the AZdesc set), including descriptors for molecular size, lipophilicity, hydrogen bonding, electrostatics, and topology were calculated with Clab, the AstraZeneca in-house descriptor engine. The details of the AZdesc descriptors are described elsewhere.[36−38] The R-group mapping and extraction was done by using a in-house software fRGS[39] to automatically strip out all R-groups based on a user supplied scaffold for the input molecular structure file. In the case of symmetrical substituents, assignments were done in such a manner that groups having the largest number of atoms were mapped in the same R-group position. Finally the R-group stripping was manually checked for consistency.

Support vector machine (SVM) is a machine learning method commonly used to build QSAR models.[40,41] In SVM learning[42,43] a hyperplane is constructed, which discriminates between the data points of distinct classes (binary SVM) in such a way that the margin between the classes is maximized. The final position and orientation of the hyperplane is defined by a subset of training vectors, the so-called support vectors. In the current study, the epsilon-SVR algorithm implemented in the open source program LIBSVM[44] was used to build SVM models. The commonly used radial basis function kernel was employed during model building. Optimal Gamma and C parameters obtained from a grid search for the highest cross-validation accuracy were utilized in the final SVM models.

The Free-Wilson modeling was carried out using the Fit Model function of the multilinear regression module in JMP,[45] which in essence is the same as the original Free-Wilson method. The only difference is the different definition of symmetry equation, which results in different intercept and coefficients of the multilinear regression equation, but the predicted values are the same. To compare with the Free-Wilson method, we combined the LIBSVM and R-group signature descriptors to build local models. In the current study, the correlation coefficient R^2, RMSE, and the cross-validated correlation coefficient Q^2 obtained by a 10-fold cross-validation process were assessed for all the models, and all these statistical analysis were performed in JMP.

It is also interesting to see how models based on physicochemical properties perform on the selected data sets. We therefore used bulky descriptors of the full compound to build QSAR models, and due to the inability of LIBSVM to handle missing values and distinguish different descriptor value types another SVM package implemented in the in-house machine learning package AZOrange[46] was used (referred to as AZO_SVM) as the modeling algorithm for convenience. QSAR models were built by using the AZO_SVM method with the AZdesc descriptor set and compared with Free-Wilson models and other fingerprint based models.

Signature and R-Group Contributions. Any QSAR model is an approximation of the relationship between biological activity and compound characteristics (descriptors) and can therefore be viewed as a mathematical function. Different machine learning methods have different ways of deriving these approximations, but they can always be described by Taylor series expansions. The gradient of a function in a point tells how this function behaves (steepness of increase or decrease) in the neighborhood to that point, and it can be calculated for any sufficiently smooth function; this corresponds to the zero- and first-order terms in a Taylor series. Each gradient component of a QSAR model indicates how significant an impact the descriptor component has on the bioactivity prediction. Hence, it is a measure of the contribution/importance of each descriptor in the QSAR model.

Carlsson et al.[31] have previously published on how to calculate the discrete gradient to measure the contribution of individual signature descriptors in an Ames mutagenicity SVM model. The same methodology is applied here to obtain the *j*th component of the discrete gradient for the signature SVM model in each compound according to eq 1:

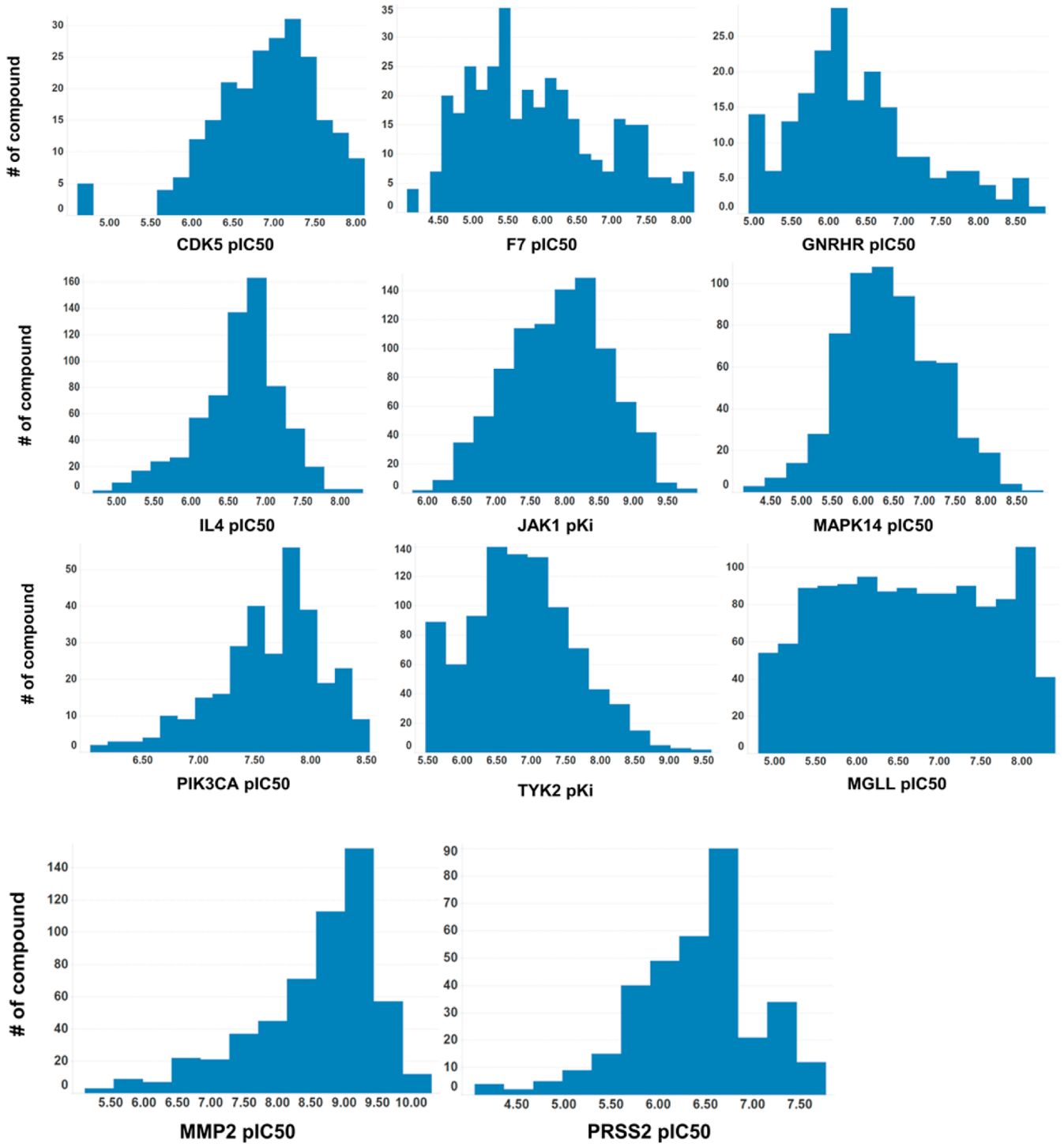


Figure 2. The distribution of bioactivity data for the used data sets.

$$\frac{Df}{Dx_j} = \frac{f(x + h_j) - f(x)}{h_j} \quad (1)$$

Here h_j is the step size of increment of the jth descriptor, x, and it is set as 1 throughout the study, and f is the QSAR function. The R-group contribution in each compound can further be calculated by adding up all gradient components corresponding to the signatures which are comprised in the R-group as shown in eq 2:

$$C_n = \sum_{j=1}^{k} t_j \times \frac{Df}{Dx_j} \quad (2)$$

Here t_j is the occurrence of the jth signature in the nth R-group of the compound, and C_n is the contribution of the R-group for one particular compound to bioactivity. It should be borne in mind that C_n does not represent the actual contribution to binding affinity, but the relative importance of the R-group and thus the R-groups can be ranked accordingly. Finally each R-group contribution for the model is an average value over the

dx.doi.org/10.1021/ci4001376 | J. Chem. Inf. Model. 2013, 53, 1324−1336

Table 2. Model Performance for the Whole Test Sets

	Mol_ECFP6_SVM[a]		Mol_Sign_SVM[b]		FW[c]		Rgp_Sign_SVM[d]		AZdesc_AZOSVM[e]	
data set	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
CDK5	0.59	0.38	0.62	0.36	0.57	0.38	0.58	0.38	0.38	0.47
F7	0.62	0.54	0.74	0.44	0.66	0.46	0.73	0.46	0.64	0.52
GNRHR	0.58	0.57	0.53	0.60	0.04	1.09	0.51	0.61	0.6	0.56
IL4	0.68	0.36	0.67	0.36	0.56	0.42	0.64	0.38	0.65	0.37
JAK1	0.5	0.49	0.59	0.44	0.81	0.32	0.62	0.42	0.35	0.56
MAPK14	0.3	0.61	0.37	0.59	0.20	0.63	0.38	0.59	0.36	0.59
MGLL	0.46	0.71	0.59	0.62	0.53	0.65	0.57	0.63	0.45	0.72
MMP2	0.55	0.68	0.65	0.59	0.97	0.16	0.66	0.59	0.23	0.88
PIK3CA	0.41	0.34	0.46	0.32	0.43	0.26	0.53	0.3	0.33	0.36
PRSS2	0.69	0.36	0.73	0.34	0.64	0.36	0.67	0.37	0.67	0.37
TYK2	0.56	0.5	0.62	0.47	0.75	0.42	0.62	0.47	0.58	0.49

[a]LIBSVM model based on ECFP6 fingerprints. [b]LIBSVM model based on molecular signatures. [c]Free-Wilson model, it cannot make prediction for the full test set. [d]LIBSVM model based on R-group signatures. [e]AZOrange SVM model based on AZdescriptor set.

Table 3. Comparison of Model Performance on Free-Wilson Predicted Test Set[a]

		Mol_ECFP6_SVM		Mol_Sign_SVM		Rgp_Sign_SVM		FW		AZdesc_AZOSVM	
data set	nr. compd	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
CDK5	44	0.61	0.37	0.62	0.36	0.62	0.36	0.57	0.38	0.38	0.46
F7	44	0.61	0.49	0.69	0.43	0.7	0.42	0.66	0.46	0.62	0.48
GNRHR	10	0.87	0.41	0.82	0.48	0.65	0.66	0.04	1.09	0.73	0.58
IL4	120	0.74	0.32	0.72	0.33	0.72	0.34	0.56	0.42	0.74	0.32
JAK1	66	0.57	0.48	0.7	0.4	0.77	0.35	0.81	0.32	0.48	0.53
MAPK14	17	0.22	0.62	0.3	0.59	0.2	0.62	0.2	0.63	0.24	0.61
MGLL	115	0.52	0.65	0.65	0.56	0.6	0.6	0.53	0.65	0.4	0.73
MMP2	24	0.33	0.77	0.9	0.3	0.94	0.23	0.97	0.16	0.08	0.9
PIK3CA	42	0.53	0.24	0.57	0.23	0.57	0.23	0.43	0.26	0.3	0.29
PRSS2	38	0.76	0.3	0.77	0.29	0.72	0.32	0.64	0.36	0.61	0.38
TYK2	70	0.67	0.49	0.74	0.44	0.76	0.42	0.75	0.42	0.61	0.53

[a]The model names are the same as for Table 2.

Table 4. 10-Fold Cross-Validation Results for Different Modeling Strategies[a]

	Mol_ECFP6_SVM		Mol_Sign_SVM		Rgp_Sign_SVM		AZdesc_AZOSVM	
data set	Q^2	RMSE	Q^2	RMSE	Q^2	RMSE	Q^2	RMSE
CDK5	0.54	0.45	0.62	0.41	0.58	0.44	0.58	0.36
F7	0.73	0.50	0.75	0.48	0.74	0.50	0.70	0.53
GNRHR	0.49	0.62	0.47	0.64	0.44	0.65	0.46	0.64
IL4	0.64	0.34	0.63	0.34	0.63	0.34	0.58	0.36
JAK1	0.58	0.47	0.58	0.47	0.63	0.43	0.50	0.51
MAPK14	0.34	0.64	0.31	0.66	0.37	0.62	0.28	0.66
MGLL	0.51	0.69	0.60	0.62	0.54	0.66	0.49	0.70
MMP2	0.61	0.59	0.67	0.55	0.70	0.53	0.55	0.64
PIK3CA	0.37	0.38	0.40	0.38	0.47	0.35	0.27	0.41
PRSS2	0.60	0.42	0.65	0.39	0.60	0.42	0.61	0.42
TYK2	0.60	0.51	0.64	0.47	0.68	0.44	0.61	0.49

[a]The model names are the same as for Table 2.

whole training set. To validate the R-group contribution generated from the SVM model gradient, a linear regression analysis was carried out to relate the gradient based R-group contributions with the R-group contributions from the Free-Wilson analysis.

RESULTS AND DISCUSSIONS

The primary objective of this article is to present a novel R-group QSAR methodology that overcomes the drawback of the Free-Wilson method on the scope of predictions without sacrificing the model interpretability. The eleven public data sets which were chosen comprise a variety of chemical scaffolds and biological targets. Furthermore, the size of the data sets is also sufficiently large to allow for solid statistical comparison between the chosen QSAR methods.

Model Performance. The distribution of bioactivity data for the eleven data sets is shown in Figure 2. Several types of models were built for each data set, including models built on ECFP6 fingerprints (with 1024 bits), molecular signatures, signatures for R-groups only, and conventional physicochemical descriptors. The ECFP6 fingerprint with 2048 bits was also

dx.doi.org/10.1021/ci4001376 | J. Chem. Inf. Model. 2013, 53, 1324−1336

Table 5. Linear Regression Results between R-Group Contributions from SVM Models and That of Free-Wilson Analysis

data set	nr. compd	nr. R-groups	R^2	RMSE
F7	292	204	0.62	0.47
JAK1	736	573	0.50	0.47
TYK2	736	576	0.71	0.32
CDK5	184	53	0.33	0.29
GNRHR	159	157	0.005	0.44
IL4	532	241 (11[a])	0.42	0.38
MAPK14	488	476	0.0007	0.94
MGLL	982	681 (3[a])	0.35	0.79
MMP2	439	473	0.42	0.66
PIK3CA	243	122	0.52	0.32
PRSS2	271	166	0.64	0.37

[a]Outliers which were excluded in the regression analysis.

tested, and the model performance was found to be very similar to the models using 1024 bits; therefore, only results of 1024 bits are reported. The model performance on all compounds in the external test sets and on the subsets of the test sets which were predicted by the Free-Wilson models are shown in Tables 2 and 3, respectively. To check the robustness of the modeling strategies, a 10-fold cross-validation process was done on all the SVM models, and the results are shown in Table 4. It can be seen that for all the SVM models, the 10-fold cross-validation Q^2 is generally at the same level with that of the randomly picked test set. This means that the randomly selected test sets do not bring in any specific bias.

In most cases the model performance drops from the Free-Wilson subset to the complete test set independent of the descriptor set. The R^2 decreases in nine out of eleven R-group SVM models by 5−30% for the complete test set compared to the Free-Wilson subset. The largest drop is for the MMP2 model where R^2 decreases from 0.94 to 0.66. The latter is in line with the Q^2 (0.7) from the 10-fold cross-validation, which may indicate that the structural information of the Free-Wilson subset is better covered in the training set and results in a deceptively high correlation. A similar scenario is seen for both the molecular fingerprints and ECFPs. In F7 and MAPK14 models, comparing with the Free-Wilson subset, the R^2

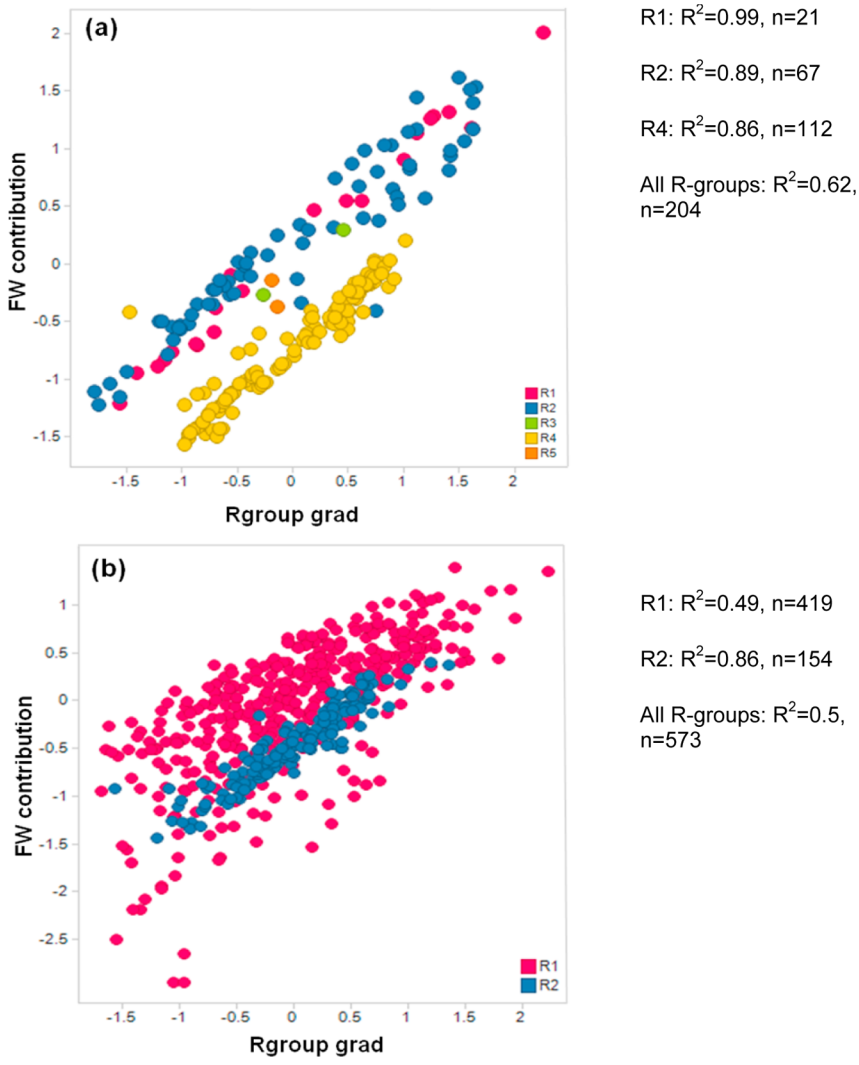


Figure 3. The correlations between R-group contributions from Free-Wilson models and R-group signature models for (a) F7 and (b) JAK1 data sets. The *y* axis shows the R-group contribution of the Free-Wilson model, and the *x* axis shows the R-group contribution of the signature model. Data points are colored according to their R-group positions. The regression analyses were done for each individual R-group position and all R-groups, respectively.

dx.doi.org/10.1021/ci4001376 | J. Chem. Inf. Model. 2013, 53, 1324−1336

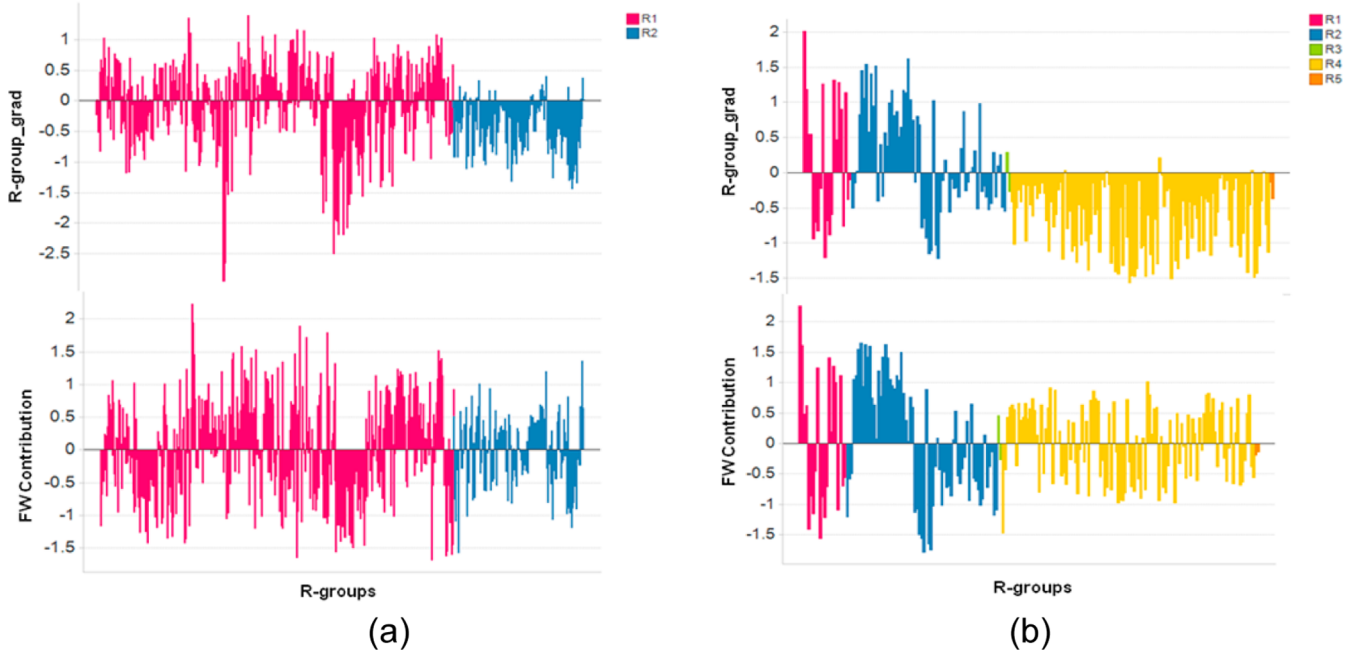


Figure 4. Comparison of R-group contributions in the SVM models and the Free-Wilson models for (a) JAK1 and (b) F7 data sets. In each plot the *x* axis corresponds to various R-groups, and the *y* axis refers to their contributions in the model. The R-groups are colored according to their substituent position on the core structure.

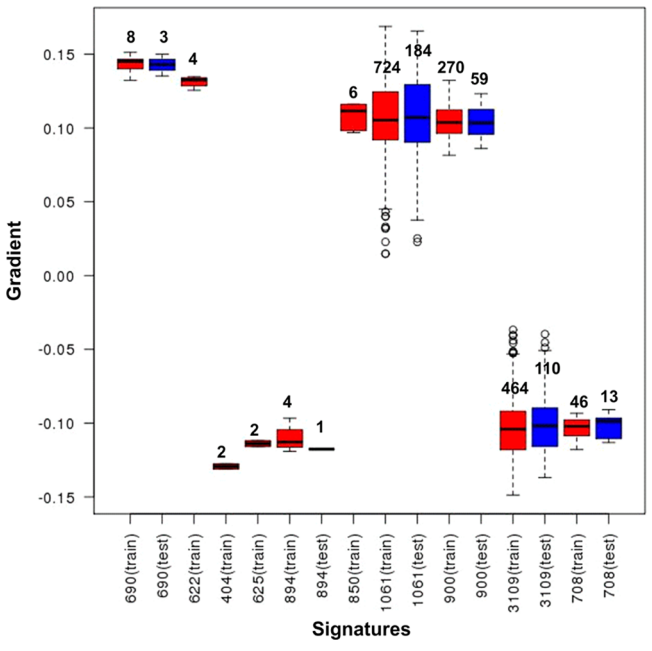


Figure 5. The box plot of gradients for the top 10 signatures in the JAK1 SVM model. The *y* axis shows the model gradient value, and the *x* axis shows the signature ID which the gradient refers to. The text in the bracket of the *x* axis label indicates if the signature is for the training or test set. The band near the middle of the box is the median value of the gradient distribution. The bottom and top of the box refer to lower and upper quartiles, the bottom and upper whisker correspond to lower and upper adjacent value, and the circles refer to outliers. The red boxes are meant for signatures from the training set and blue boxes for the test set. The number above the box refers to the occurring frequency of the signature in the compound set.

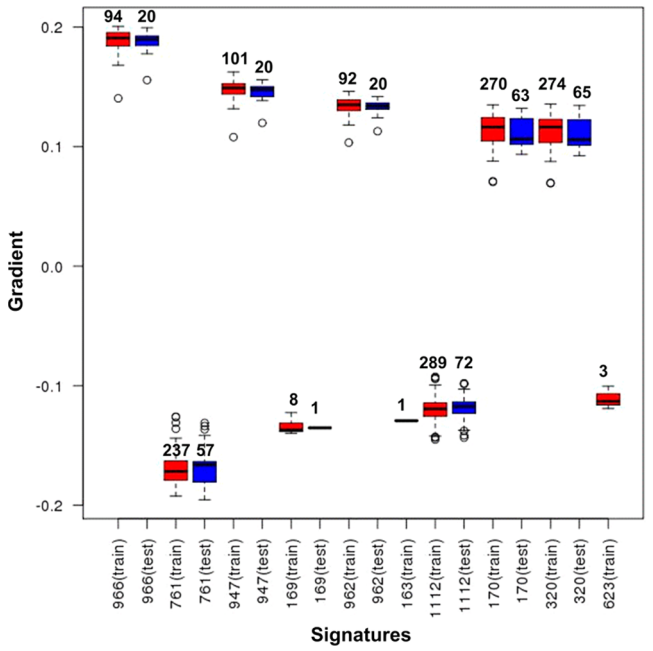
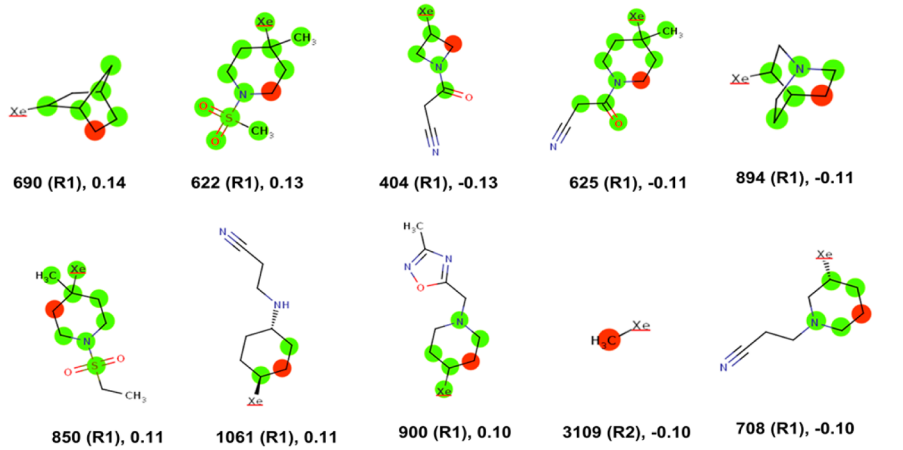


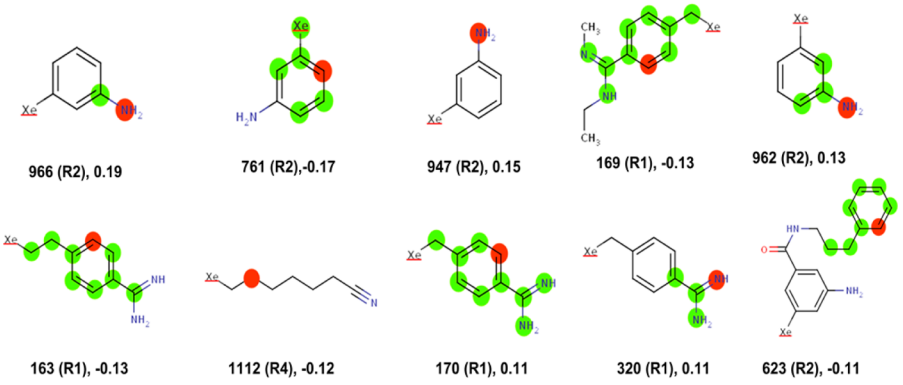
Figure 6. The box plot of gradients for the top 10 signatures in the F7 SVM model. The meaning of the labels and legend are the same as for Figure 5.

increases for the full test set. For the MAPK14 set the R-group SVM model performs rather poorly, and the increase of prediction R^2 from 0.2 to 0.38 merely reflects the incomplete statistics of the Free-Wilson subset which only covers 17 out of the 122 compounds in the full test set. Also in this case the Q^2 (0.37) of the 10-fold cross-validation is in line with that of the full external test set. The R^2 of the F7 model has a modest improvement for the full test set and is in agreement with the Q^2 in 10-fold cross-validation for the training set. The results are similar for the ECFPs and molecular signature models. Further analysis on the Free-Wilson test set results reveals that the molecular signature model performs best or equally best on 5 data sets followed by the R-group signatures and ECFPs

dx.doi.org/10.1021/ci4001376 | J. Chem. Inf. Model. 2013, 53, 1324−1336



(a)



(b)

Figure 7. Top 10 R-group signatures for (a) JAK1 and (b) F7 SVM models. Each signature is mapped on an example R-group existing in the training set. In the label text of each structure, the signature ID is listed at the beginning, the text in the bracket refers to the substituent position which the R-group belongs to, and the floating number is the average gradient value for the signature in the training set. The red atom refers to the center atom of the signature, and green atoms correspond to the neighboring atoms which are within three bond lengths to the center atom. The Xe atom here refers to the attachment point for the R-group.

being best or equally best on 4 and 2 targets, respectively. Free-Wilson models perform best on two data sets, and the AZdesc model performs equally best with ECFPs on the IL4 data set. The picture remains quite similar when analyzing the performance on the full test set. The full compound and R-group signature models perform best or equally best on 5 targets, respectively. The ECFPs is the best performing model on the IL4 set, whereas the AZdesc model performs best on the GNRHR set. In summary, four observations can be made: (i) It seems that the models built on fingerprints (molecular signatures, R-group signatures, and ECFP6 fingerprints) generally have similar quality in terms of prediction accuracy on the external test sets. (ii) Physicochemical property-based models generally perform worse than that of fingerprint-based models, and it is probably due to the fact that fingerprints can catch some subtle chemical functional groups required for ligand binding which are not encoded in physicochemical property-based bulky descriptors. (iii) Signature models seem, in most cases, to be slightly better than the models using ECFPs. (iv) The molecular signature models seem to perform slightly better than the R-group signatures which may indicate that some interaction terms with the scaffold are important. However, in most cases the R-group model is of comparable quality. This implies that the full compound fingerprint already comprise the information encoded in R-group signatures, but from a model interpretation perspective, the R-group signature models have significant advantage and will be preferred since the influential signature bit for the model can be traced to detailed substitution positions. This is impossible to identify in full compound models if there are certain functional groups which occur in the core of the structure part and multiple R-group positions simultaneously.

Provided that bioactivity data are linearly related to the R-groups in the Free-Wilson analysis and the fingerprints in the SVM model, both models should in theory perform similarly. As we have discussed, Free-Wilson models can only make a prediction on a subset of the full test set, and their performances are listed in Table 3. The prediction results of all other models on the same test set are also given in Table 3, and it can be seen that among the 11 data sets, only in two cases (JAK1 and MMP2 data sets) Free-Wilson models perform best, whereas in all other cases, their performances are worse than fingerprint based methods. If we take a closer look at the JAK1 and MMP2 cases, the performance of R-group signature models for JAK1 (R^2 = 0.77) is actually comparable to that of Free-Wilson models (R^2 = 0.81), and for MMP2 it is

dx.doi.org/10.1021/ci4001376 | J. Chem. Inf. Model. 2013, 53, 1324−1336

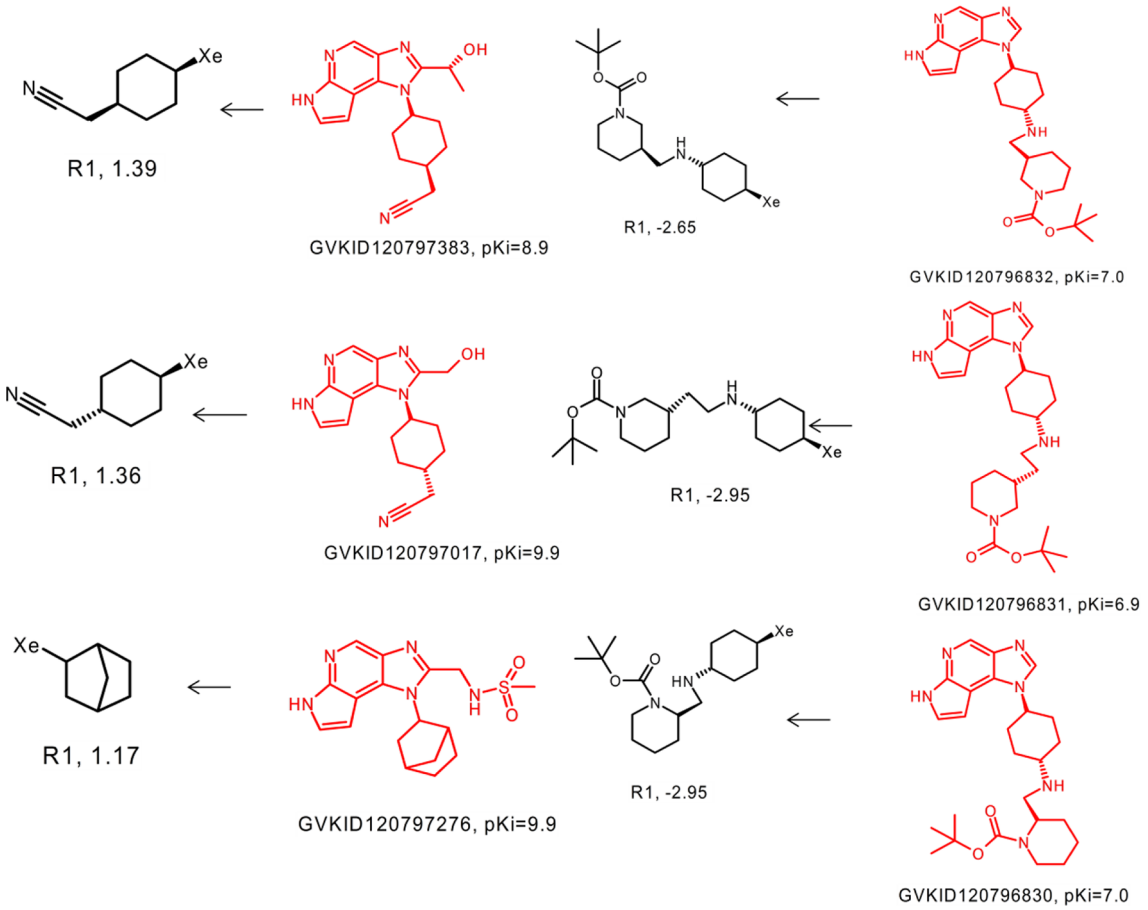


Figure 8. The top three R-groups which contribute positively (left) and negatively (right) to bioactivity in the JAK1 signature SVM model. The R-group position and their contribution values are listed. The parent compounds which comprise these R-groups are colored in red.

0.94 and 0.97, respectively. From Table 3, it can also be seen that in most cases the models with bulky descriptors based on molecular physicochemical properties performed worse than the fingerprint models. The GNRHR models are a striking exception, where all SVM models (including fingerprint and bulky descriptor models) perform well (R^2 = 0.65−0.87); the Free-Wilson does not provide a meaningful model (R^2 = 0.04). This indicates that there may be a nonlinear relationship which the Free-Wilson method cannot handle. Overall we can see that fingerprint based models (especially for the R-group signature model) perform either better than or comparable to Free-Wilson models and at the same time have the advantage of being able to predict beyond the training set.

R-Group Contributions of SVM Models. As we have emphasized, one major benefit of Free-Wilson analysis is that the model can be interpreted by examining the contribution for each R-group, which provides clear and straightforward guidance for the medicinal chemist to design new compounds. Nonlinear models performs well in predictions but are often flawed by its lack of interpretability. To overcome this we introduce the concept of the SVM model gradient. The R-group contribution in our method can also be calculated based on eqs 1 and 2, and should provide similar interpretability as the Free-Wilson model. A direct way of validating our R-group contribution is to make a comparison with the R-group contribution from the Free-Wilson model to investigate if there is a significant correlation between these two metrics. Admittedly the R-group contributions in a nonlinear SVM model are *per se* not equivalent to the linear contributions of the Free-Wilson model. However if both models perform equally well, it strongly indicates that a dominant linear relationship exists in the data set. Thus, our hypothesis is that statistically good models imply good correlations between Free-Wilson and R-group signature contributions. For all the R-group signature models, R-group contributions were calculated according to our previous description, and Free-Wilson R-group contributions were generated with the JMP software. A linear regression was then carried out to relate both contribution metrics, and the results are shown in the Table 5 (detailed scatter plots are displayed in Figure S1 of the Supporting Information). For IL4 and MGLL data sets, 11 and 3 outlier R-groups were removed, respectively, before the linear regression analysis was performed. Those outliers all have a quite large Free-Wilson R-group contribution (absolute value larger than 2.2) comparing with the other R-groups. Much to our satisfaction, the plots reveal that significant correlations (R^2 ranging from 0.33 to 0.71) can be seen between two contribution metrics in all but the GNRHR and MAPK14 sets. For TYK2, F7, and PRSS2 data sets, good correlations (R^2 > 0.6) are obtained. These results show that the R-group contributions calculated from SVM model gradients do correlate with the bioactivity data and can be used as a relative measure of R-group importance. The low correlation in the MAPK14 set is understandable as both the Free-Wilson and the R-group SVM model perform poorly (R^2 ∼ 0.2) on this data set. Similarly the lack of correlation in the GNRHR set is a

dx.doi.org/10.1021/ci4001376 | J. Chem. Inf. Model. 2013, 53, 1324−1336

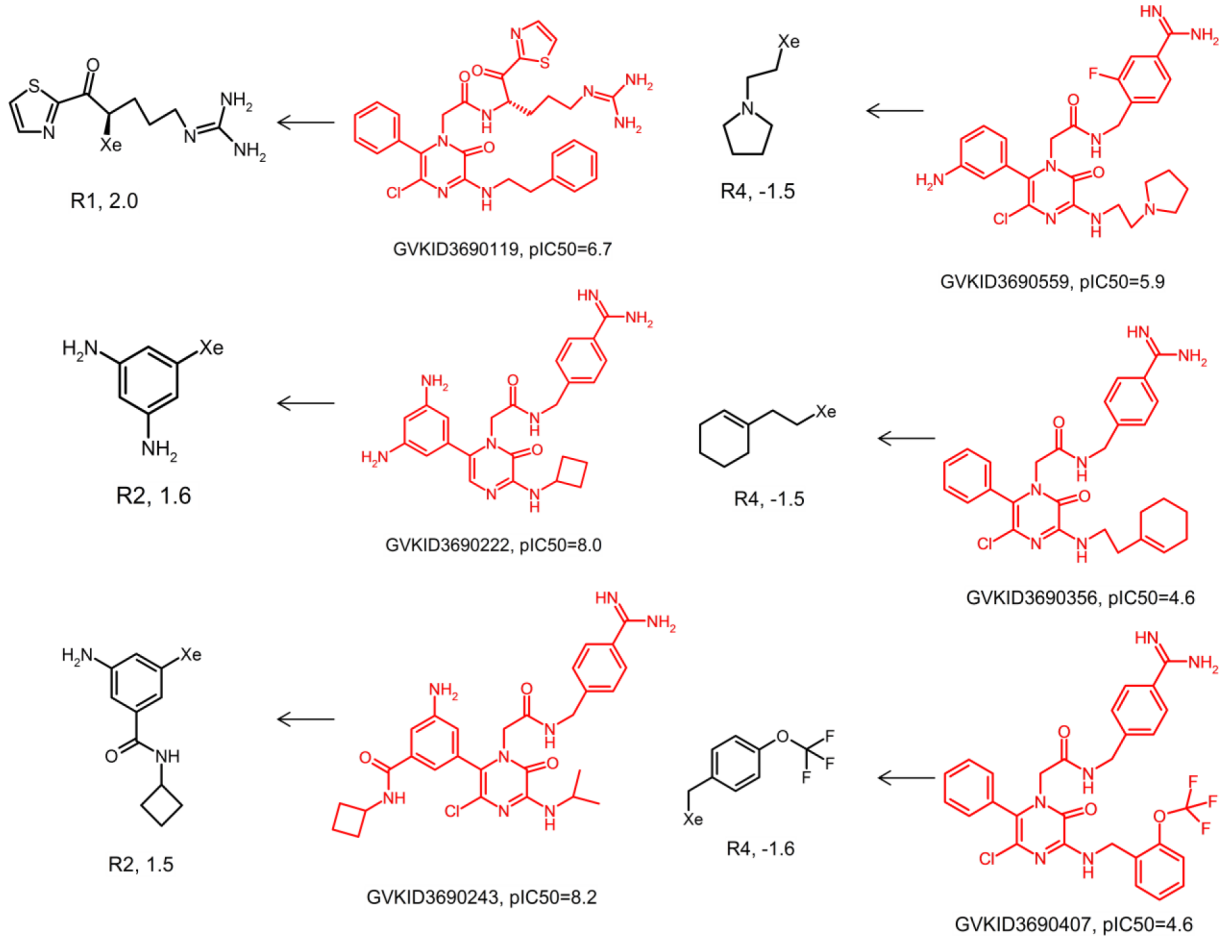


Figure 9. The top three R-groups which contribute positively (left) and negatively (right) to bioactivity respectively in the F7 signature SVM model. The R-group position and their contribution values are listed. The parent compounds which comprise these R-groups are colored in red.

consequence of the absence of a predictive Free-Wilson model for this target and may suggest a nonlinear structure−activity relationship picked up by the SVM model. The F7 and JAK1 data sets were chosen for a more detailed analysis on the correlation of the R-group contributions (shown in Figure 3a,b). The data in Figure 3 are color coded according to the position of R-groups. For both data sets the R-groups are split into different subgroups according to their substitute positions on the core structure (shown in Figure 1), and a separate regression analysis was done to each subgroup. For the F7 set, the R3, R5 subgroups were skipped due to too few data. In Figure 3a the correlations of R1 (red), R2 (blue), and R4 (yellow) contributions for the F7 data set are quite spectacular, and very high correlations are obtained ($R^2 > 0.8$). When all the R-groups are included in the regression analysis, the R^2 drops down to 0.62. In the JAK1 set the correlation coefficient R^2 for R1 (red) and R2 (blue) subgroups is 0.49 and 0.89, respectively (shown in Figure 3b), while R^2 for all R1 and R2 substituents is 0.5. Although it is unclear why R^2 of R1 and R2 subgroups have such a large difference, the correlation for the R2 subgroup alone is obviously much better than mixing up R1 and R2 subgroups. Thus, it may suggest that the correlations of R-group contributions between SVM and Free-Wilson models could be higher when looking into each substitute position.

In Figure 4 the R-group contributions from SVM and Free-Wilson models are aligned side by side for comparison. Notably, the distribution pattern of R2 contribution in the JAK1 SVM model and R4 contribution in the F7 SVM model are significantly different from that of corresponding Free-Wilson models, whereas the other R-groups align well. Most of the SVM R-group contributions in these two subgroups (R2 in the JAK1 set and R4 in the F7 set) are negatively shifted comparing with Free-Wilson models. However, from Figure 3, we know that these R-group contributions are highly correlated between the SVM and Free-Wilson models. This implies that the SVM R-group contribution value does not reflect the absolute contribution to bioactivity but a relative ranking of R-group's contribution instead, and this ranking should better be applied to R-groups belonging to the same substituent position, while the Free-Wilson R-group contribution more or less represents the absolute bioactivity contribution of individual R-groups. This is actually understandable because the R-group contributions in Free-Wilson analysis are obtained by linearly fitting to the bioactivity data, whereas the R-group contributions in the SVM model are calculated based on model gradients which are not directly related with bioactivity data.

Model Interpretability for JAK1 and F7 Data Sets. Another feature of the R-group signature model is that the contribution can be further decomposed to the signature level as described in eq 5. Here the JAK1 and F7 data sets are chosen as examples to demonstrate this. The box plots in Figures 5 and 6 show the distribution of the top 10 signatures in terms of gradient absolute value for JAK1 and F7 data sets, respectively. Each plot includes the five signatures making the largest

dx.doi.org/10.1021/ci4001376 | J. Chem. Inf. Model. 2013, 53, 1324−1336

positive contribution to the model and also the top five signatures contributing most negatively. The detailed structures of those signatures are shown in Figure 7. In Figure 5, the gradients for signature no. 1061 and 3109 have much larger variation than other signatures, and they represent small fragments which occur frequently in the training set (724 and 464 times). This means that the contribution for these two signatures depend heavily on their surrounding environment. While signature no. 900 also occurs frequently in the training set, its contribution varies much less than signature no. 1061 and 3109. For the F7 data set (Figure 6), all frequently occurring signatures seem to have much less variation compared with signature no. 1061 and 3109, respectively, in the JAK1 training set. The gradient distributions of those signatures which exist in both training and test sets are also compared, and it appears that the distributions for the same signature in both the training and test set are fairly similar. Figure 7 shows the average gradient values for these top 10 signatures in JAK1 and F7 data sets, respectively, and they should, in principle, represent the contributions of these signatures to the bioactivity. Comparing signature no. 622 and 625 in Figure 7a, it seems that these two signatures are quite similar, but changing the sulfonamide group to amide decreases the contribution. The gradients values for signature no. 900 and 708 imply that changing the position of the attachment point on the piperidine ring of the R1 position from the meta position to the para position may have a positive influence on the bioactivity. The positive gradient values for signature no. 966, 947, and 962 in Figure 7b suggest that adding an amino group on the R2 phenyl ring could have a positive effect on bioactivity. It seems that examining the bioactivity contribution at the signature level can also provide some useful SAR information, but it should always be borne in mind that the gradient value listed in the figure only corresponds to the signatures which are mapped on the R-group structure in the figure and does not mean the whole R-group contribution. The R-group contribution is calculated via summing up the gradient values of all the signatures contained in the R-group. So when new R-groups based on the signature contributions are designed, the R-group contribution should always be calculated to make sure the final additive effect of the R-group is positive.

The top three R-groups contributing positively and negatively for JAK1 and F7 sets are displayed in Figures 8 and 9, respectively. For each R-group, one parent compound structure is also displayed together with its bioactivity data for clarification. For JAK1 inhibition, three cyclohexane based R1 substitutions have the largest contribution, which is probably due to the contribution of hydrophobic signatures such as no. 690 and 1061. The three most negative R-group contributions are related to three large R1 substituents, and compounds containing these R-groups generally have low activity. For the F7 set, a guanidine containing R1 group is ranked as the group which has the largest positive contribution to F7 inhibition. The second and third positive R2 groups have an amino substituted phenyl ring, and this is consistent with the signature contributions shown in Figure 7b. The three most negatively contributing groups are all R4 substituents. In the top 10 signatures in Figure 7b, there is only one negative signature which comes from the R4 substituent. Hence, the negative contributions of those R4 substituents are probably due to some other negatively impacted signatures. Actually if the examined signature list is extended to the top 20, there will be two more negative R4 signatures showing up. Inspecting the whole data set also shows that compounds having those R4 groups generally have fairly low activity. It is worthwhile to mention that the current handling of small signatures containing only one or two atoms is still not accurate enough. These small signatures appear frequently in compounds, and their contribution to bioactivity could be different depending on, for example, how far they are away from the core structure. One direction for the future development would be to encode this distance information into signature to further improve the model, and we will report the progress on this aspect in due course.

Overall these R-group contribution analysis results show that the nonlinear R-group SVM model provides similar interpretability as the Free-Wilson analysis and can be an intuitive tool for medicinal chemists to design new compounds. Moreover, the R-group SVM models have in general better prediction accuracy than Free-Wilson models, and by employing the R-group signature our new methodology has the superiority in increasing the prediction domain of the test set, which is a key limitation of Free-Wilson analysis. We may also conclude that the R-group signature SVM methodology compares well to the molecular FPs (ECFP and signatures) presented in this work as well as to the models using the traditional bulk descriptors. Admittedly, we have not performed an extensive evaluation using other QSAR methodologies, as this has not been the purpose of the paper, but we are quite confident that our approach will perform well in comparison with these methods. Finally it must be emphasized that this methodology is not necessarily restricted to signature descriptors, and other types of structural fingerprints generated from R-groups can also be expected to be applicable to this method.

CONCLUSIONS

Free-Wilson analysis has obvious advantages in local QSAR models; for example, model interpretability and no need for substituent parameters.[13] However, its application is largely restrained by the fact that it is not applicable for making predictions outside the scope defined by the already used training set R-groups. Here we introduce a novel methodology to overcome these drawbacks without sacrificing the model interpretability. We have used the R-group signatures as descriptors to build nonlinear SVM models. Because of the more general description of signatures, our method can make predictions on compounds containing R-groups which are not present in the training set. The results for the nonlinear SVM models using either molecular or R-group fingerprints are, in most cases, as good as or better than the corresponding Free-Wilson models. Most importantly, the concept of the SVM model gradient is introduced in the current study as a measure of individual signature as well as whole R-group model contributions. The strong correlation between R-group contributions of the Free-Wilson and the signature SVM models indicates a dominant linear SAR as requested by the Free-Wilson method and forms the basis for the comparison of contribution from R-group signatures and Free-Wilson R-groups. Our results show that in the majority of the eleven selected focused data sets, R-group contributions derived from SVM models demonstrate a significant correlation with Free-Wilson R-group coefficients. These results highlight that gradient-based R-group contributions provide clear interpretability to nonlinear SVM models and therefore have the potential to become a powerful modeling tool with applications in any drug discovery project.

■ ASSOCIATED CONTENT

Supporting Information

The detailed scatter plots for correlations between R-group contributions from SVM and Free-Wilson models and all eleven data sets used in the current study including chemical structures and bioactivity data. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: hongming.chen@astrazeneca.com (H.C.), ingemar.nilsson@astrazeneca.com (I.N.).

Present Address

[&]Department of Computational Chemistry, H. Lundbeck A/S, Ottiliavej 9, 2500 Valby, Denmark.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors would like to thank Dr. Ola Engkvist, Dr. John Cumming, and Dr. Willem Nissink for valuable discussions and Dr. Christian Tyrchan, Dr. Jonas Boström, and Dr. Allen Plowright for critical revisions of the manuscript.

■ REFERENCES

(1) Paul, S. M.; Mytelka, D. S.; Dunwiddie, C. T.; Persinger, C. C.; Munos, B. H.; Lindborg, S. R.; Schacht, A. L. How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nat. Rev. Drug Discovery* **2010**, *9*, 203−14.

(2) Plowright, A. T.; Johnstone, C.; Kihlberg, J.; Pettersson, J.; Robb, G.; Thompson, R. A. Hypothesis driven drug design: improving quality and effectiveness of the design-make-test-analyse cycle. *Drug Discovery Today* **2012**, *17*, 56−62.

(3) Andersson, S.; Armstrong, A.; Björe, A.; Bowker, S.; Chapman, S.; Davies, R.; Donald, C.; Egner, B.; Elebring, T.; Holmqvist, S.; Inghardt, T.; Johannesson, P.; Johansson, M.; Johnstone, C.; Kemmitt, P.; Kihlberg, J.; Korsgren, P.; Lemurell, M.; Moore, J.; Pettersson, J. A.; Pointon, H.; Pontén, F.; Schofield, P.; Selmi, N.; Whittamore, P. Making medicinal chemistry more effective–application of Lean Sigma to improve processes, speed and quality. *Drug Discovery Today* **2009**, *14*, 598−604.

(4) Hutter, M. C. In silico prediction of drug properties. *Curr. Med. Chem.* **2009**, *16*, 189−202.

(5) Cramer, R. D.; Patterson, D. E.; Bunce, J. D. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959−67.

(6) Li, H.; Yap, C. W.; Xue, Y.; Li, Z. R.; Ung, C. Y.; Han, L. Y.; Chen, Y. Z. Statistical learning approach for predicting specific pharmacodynamic, pharmacokinetic, or toxicological properties of pharmaceutical agents. *Drug Dev. Res.* **2005**, *66*, 245−259.

(7) Chadwick, A.; Hajek, M. Learning to improve the decision-making process in research. *Drug Discovery Today* **2004**, *9*, 251−257.

(8) Duch, W.; Swaminathan, K.; Meller, J. Artificial intelligence approaches for rational drug design and discovery. *Curr. Pharm. Des.* **2007**, *13*, 1497−1508.

(9) Yap, C.; Xue, Y.; Li, H.; Li, Z.; Ung, C.; Han, L.; Zheng, C.; Cao, Z.; Chen, Y. Prediction of compounds with specific pharmacodynamic, pharmacokinetic or toxicological property by statistical learning methods. *Mini-Rev. Med.Chem.* **2006**, *6*, 449−459.

(10) Yap, C. W.; Li, H.; Ji, Z. L.; Chen, Y. Z. Regression methods for developing QSAR and QSPR models to predict compounds of specific pharmacodynamic, pharmacokinetic and toxicological properties. *Mini-Reviews in Med. Chem.* **2007**, *7*, 1097−107.

(11) Mager, D. E. Quantitative structure-pharmacokinetic/pharmacodynamic relationships. *Adv. Drug Delivery Rev.* **2006**, *58*, 1326−56.

(12) Free, S. M.; Wilson, J. W. A mathematical contribution to structure-activity studies. *J. Med. Chem.* **1964**, *7*, 395−9.

(13) Craig, P. N. Proceedings: comparison of Hansch and free-Wilson methods for structure-activity correlation. *Cancer Chemother. Rep., Part 2* **1974**, *4*, 39.

(14) Kubinyi, H.; Kehrhahn, O. H. Quantitative structure-activity relationships. 3.1. A comparison of different Free-Wilson models. *J. Med. Chem.* **1976**, *19*, 1040−9.

(15) Hansch, C.; Muir, R. M.; Fujita, T.; Maloney, P. P.; Geiger, F.; Streich, M. The correlation of biological activity of plant growth regulators and chloromycetin derivatives with Hammett constants and partition coefficients. *J. Am. Chem. Soc.* **1963**, *719*, 2817−24.

(16) Hansch, C.; Maloney, P. P.; Fujita, T.; Muir, R. M. Correlation of biological activity of phenoxyacetic acids with Hammett substituent constants and partition coefficients. *Nature* **1962**, *194*, 178−180.

(17) Nilsson, I.; Polla, M. O. Composite multi-parameter ranking of real and virtual compounds for design of MC4R agonists: renaissance of the Free-Wilson methodology. *J. Comput.-Aided Mol. Des.* **2012**, *26*, 1143−57.

(18) Goldberg, F. W.; Leach, A. G.; Scott, J. S.; Snelson, W. L.; Groombridge, S. D.; Donald, C. S.; Bennett, S. N. L.; Bodin, C.; Gutierrez, P. M.; Gyte, A. C. Free-Wilson and structural approaches to co-optimizing human and rodent isoform potency for 11β-hydroxysteroid dehydrogenase type 1 (11β-HSD1) inhibitors. *J. Med. Chem.* **2012**, *55*, 10652−61.

(19) Jorissen, R. N.; Reddy, G. S. K. K.; Ali, A.; Altman, M. D.; Chellappan, S.; Anjum, S. G.; Tidor, B.; Schiffer, C. A.; Rana, T. M.; Gilson, M. K. Additivity in the analysis and design of HIV protease inhibitors. *J. Med. Chem.* **2009**, *52*, 737−54.

(20) Sciabola, S.; Stanton, R. V; Johnson, T. L.; Xi, H. Application of Free-Wilson selectivity analysis for combinatorial library design. *Methods Mol. Biol.* **2011**, *685*, 91−109.

(21) Höfgen, N.; Stange, H.; Schindler, R.; Lankau, H.-J.; Grunwald, C.; Langen, B.; Egerland, U.; Tremmel, P.; Pangalos, M. N.; Marquis, K. L.; Hage, T.; Harrison, B. L.; Malamas, M. S.; Brandon, N. J.; Kronbach, T. Discovery of imidazo[1,5-a]pyrido[3,2-e]pyrazines as a new class of phosphodiesterase 10A inhibitiors. *J. Med. Chem.* **2010**, *53*, 4399−411.

(22) Patel, Y.; Gillet, V. J.; Howe, T.; Pastor, J.; Oyarzabal, J.; Willett, P. Assessment of additive/nonadditive effects in structure-activity relationships: implications for iterative drug design. *J. Med. Chem.* **2008**, *51*, 7552−62.

(23) Tomic, S.; Nilsson, L.; Wade, R. C. Nuclear receptor-DNA binding specificity: a COMBINE and Free-Wilson QSAR analysis. *J. Med. Chem.* **2000**, *43*, 1780−92.

(24) Freeman-Cook, K. D.; Amor, P.; Bader, S.; Buzon, L. M.; Coffey, S. B.; Corbett, J. W.; Dirico, K. J.; Doran, S. D.; Elliott, R. L.; Esler, W.; Guzman-Perez, A.; Henegar, K. E.; Houser, J. A.; Jones, C. S.; Limberakis, C.; Loomis, K.; McPherson, K.; Murdande, S.; Nelson, K. L.; Phillion, D.; Pierce, B. S.; Song, W.; Sugarman, E.; Tapley, S.; Tu, M.; Zhao, Z. Maximizing lipophilic efficiency: the use of Free-Wilson analysis in the design of inhibitors of acetyl-CoA carboxylase. *J. Med. Chem.* **2012**, *55*, 935−42.

(25) Daylight Manual. http://www.daylight.com/dayhtml/doc/theory/theory.finger.html (accessed Jan 2, 2013).

(26) UNITY 2D fingerprint; Tripos Inc.: St. Louis, MO, USA.

(27) Accelrys Whitepaper; The keys to understanding MDL keyset technology. http://accelrys.com/products/pdf/keys-to-keyset-technology.pdf (accessed Jan 2, 2013).

(28) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−54.

(29) Faulon, J.-L.; Visco, D. P.; Pophale, R. S. The signature molecular descriptor. 1. Using extended valence sequences in QSAR and QSPR studies. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 707−20.

(30) Rosenbaum, L.; Hinselmann, G.; Jahn, A.; Zell, A. Interpreting linear support vector machine models with heat map molecule coloring. *J. Cheminf.* **2011**, *3*, 11.

(31) Carlsson, L.; Helgee, E. A.; Boyer, S. Interpretation of nonlinear QSAR models applied to Ames mutagenicity data. *J. Chem. Inf. Model.* **2009**, *49*, 2551−8.
(32) GOSTAR databases 2012; GVK Bioscieces Private Ltd.: Hyderabad, India.
(33) Perl programming language. http://www.perl.org/ (accessed Jan 20, 2013).
(34) Openbabel version 2.2.3. http://http://openbabel.org/wiki/Main_Page (accessed Jan 20, 2013).
(35) *Pipeline Pilot* version 8.5; Accelrys Inc.: San Diego, CA, USA.
(36) Paine, S. W.; Barton, P.; Bird, J.; Denton, R.; Menochet, K.; Smith, A.; Tomkinson, N. P.; Chohan, K. K. A rapid computational filter for predicting the rate of human renal clearance. *J. Mol. Graphics Modell.* **2010**, *29*, 529−37.
(37) Bruneau, P. Search for predictive generic model of aqueous solubility using Bayesian neural nets. *J. Chem. Inf. Comput. Sci. 41*, 1605−16.
(38) Katritzky, A. R.; Wang, Y.; Sild, S.; Tamm, T.; Karelson, M. QSPR studies on vapor pressure, aqueous solubility, and the prediction of water-air partition coefficients. *J. Chem. Inf. Model.* **1998**, *38*, 720−725.
(39) fRGS; Astrazeneca in-house software for SAR analysis.
(40) Yao, X. J.; Panaye, A.; Doucet, J. P.; Zhang, R. S.; Chen, H. F.; Liu, M. C.; Hu, Z. D.; Fan, B. T. Comparative study of QSAR/QSPR correlations using support vector machines, radial basis function neural networks, and multiple linear regression. *J. Chem. Inf. Comput. Sci. 44*, 1257−66.
(41) Darnag, R.; Mostapha Mazouz, E. L.; Schmitzer, A.; Villemin, D.; Jarid, A.; Cherqaoui, D. Support vector machines: development of QSAR models for predicting anti-HIV-1 activity of TIBO derivatives. *Eur. J. Med. Chem.* **2010**, *45*, 1590−7.
(42) Vapnik, V. N. *The Nature of Statistical Learning Theory*; Springer: New York, 1995.
(43) Vapnik, V. N. *Statistical Learning Theory*; John Wiley and Sons, Inc.: New York, 1998.
(44) Chang, C.-C.; Lin, C.-J. LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol.* **2011**, *2*, 1−27.
(45) *JMP*, version 10.0; SAS Institute Inc.: Cary, NC, USA.
(46) Stålring, J. C.; Carlsson, L. A.; Almeida, P.; Boyer, S. AZOrange - high performance open source machine learning for QSAR modeling in a graphical programming environment. *J. Cheminf.* **2011**, *3*, 28.