Journal of
Chemical Information
and Modeling

Article

pubs.acs.org/jcim

Beyond the Scope of Free-Wilson Analysis. 2: Can Distance Encoded R-Group Fingerprints Provide Interpretable Nonlinear Models?

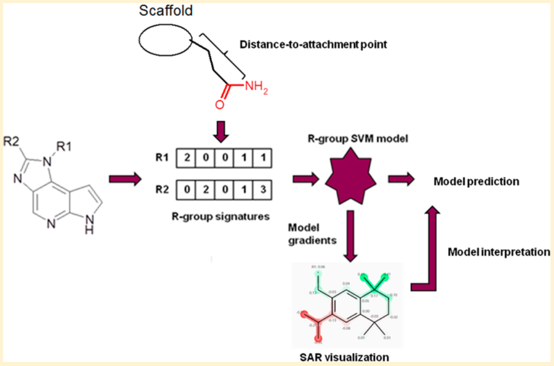
Mats Eriksson,[†] Hongming Chen,*[,†] Lars Carlsson,[§] J. Willem M. Nissink,[‖] John G. Cumming,[⊥] and Ingemar Nilsson*[,‡]

[†]Chemistry Innovation Center, Discovery Sciences, [‡]CVMD Innovative Medicines and [§]Computational Toxicology, Global Safety Assessment, AstraZeneca R&D, Mölndal 431 83, Sweden

[‖]Oncology Innovative Medicines and [⊥]Chemistry Innovation Center, Discovery Sciences, AstraZeneca R&D, Alderley Park, Macclesfield SK10 4TG, U.K.

Ⓢ *Supporting Information*

ABSTRACT: In a recent study, we presented a novel quantitative-structure–activity-relationship (QSAR) approach, combining R-group signatures and nonlinear support-vector-machines (SVM), to build interpretable local models for congeneric compound sets. Here, we outline further refinements in the fingerprint scheme for the purpose of analyzing and visualizing structure–activity relationships (SAR). The concept of distance encoded R-group signature descriptors is introduced, and we explore the influence of different signature encoding schemes on both interpretability and predictive power of the SVM models using ten public data sets. The R-group and atomic gradients provide a way to interpret SVM models and enable detailed analysis of structure–activity relationships within substituent groups. We discuss applications of the method and show how it can be used to analyze nonadditive SAR and provide intuitive and powerful SAR visualizations.



INTRODUCTION

The research and development of new small molecule drugs is a costly and lengthy endeavor.[1] The high cost of clinical trial failure has incentivized pharmaceutical companies to drive for increased success rates at all stages of the process including preclinical drug discovery. The 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. Drug designers aim to select the right compounds to synthesize from an almost infinite number of chemical structures. Quantitative structure and activity relationship (QSAR) analyses enable the building of predictive, chemical structure-based models with the potential to guide designers in an efficient way, saving time and resources.[2] Usually, a QSAR study involves the utilization of various statistical modeling and simulation techniques such as multiple linear regression (MLR),[3] partial least-squares (PLS) regression,[3,4] pattern recognition,[5] or machine learning algorithms[6−8] to predict various experimental properties *in silico*. Modeled properties include biological activity and physicochemical properties such as solubility and lipophilicity as well as various ADMET properties (e.g., clearance and oral absorption). Commonly, a predictive *in silico* model is built by employing algorithms to relate the input compound descriptors (e.g., a set of numerical values representing the 2D/3D pharmacophoric features or calculated physicochemical properties of a compound) to experimental end points. A common problem with many QSAR methods is the absence of a straightforward connection between the descriptors used for model building and the structural features of the ligands seen by the chemist. The model may be of high quality, but if the chemist cannot interpret the model in order to suggest subsequent chemical modifications, the value of the model to the chemist is diminished. Various QSAR methods based on 2D molecular fragments have been developed over the years and are frequently used in pharmaceutical research due to their simplicity and straightforward interpretation. The first fragment-based approach for QSAR analysis of a congeneric series of chemical analogues was developed in 1964 by Free and Wilson.[9] It is based on the assumption that the biological activity of a molecule can be described by a linear summation of activity contributions of its specific substructures (i.e., the parent core of the series and the corresponding substituents (R-groups)). The Free-Wilson approach uses the substituents themselves as descriptors and can therefore only be used to predict within the chemical space defined by the fragments in the training set. Although limited in its prediction scope, the Free-Wilson approach, and the modified Fujita-Ban version[10,11] have provided fruitful transparent models[12−19] for a range of experimental observations. Several methods have been developed in recent years applying molecular fingerprints for

Received: February 7, 2014
Published: March 31, 2014

dx.doi.org/10.1021/ci500075q | *J. Chem. Inf. Model.* 2014, 54, 1117−1128

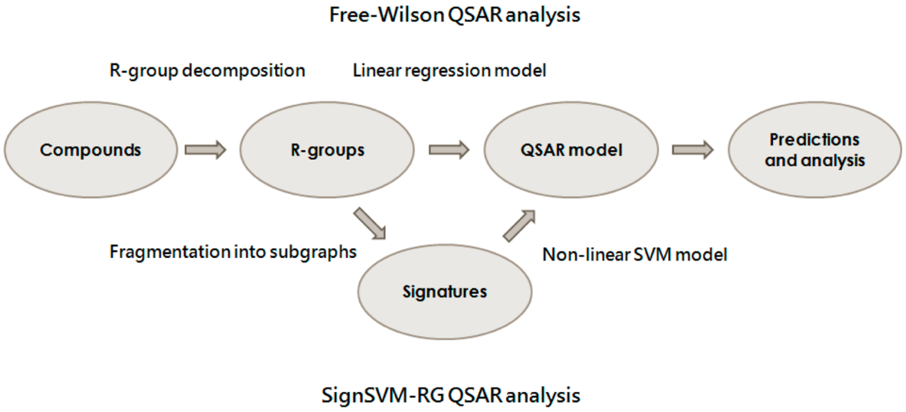


Figure 1. Workflow of Free-Wilson and SignSVM-RG QSAR analysis.

building fragment-based QSAR models.[20,21] Molecular fingerprints are representations of chemical structures originally designed to assist in chemical database substructure searching, similarity searching, nearest neighbor analysis, clustering, and classifications. There are numerous ways of representing a structure in terms of fingerprints, such as Daylight,[22] Unity,[23] MDL,[24] and ECFP/FCFP[25] commercial fingerprints. It is worth noting that there are also freely available fingerprint descriptors, such as the signature descriptor developed by Faulon et al.[26] The signature contains a systematic encoding system of atom and bond types, and, similarly to ECFP fingerprints, it describes the extended valence (i.e., circular neighborhood) of the atoms of a molecule.

We recently published a new method combining the advantages of the Free-Wilson approach and bitmap fingerprint-based QSAR methods, allowing for SAR analysis within the substituents at a given core attachment point and at the same time extending the prediction scope and improving the predictive power of the model.[27] The R-group signature SVM (SignSVM-RG) method is based on R-group decomposition data, like the Free-Wilson method, but adds another layer of information to the R-groups using signature descriptors and uses nonlinear SVM to build models. Results from this study[27] suggest that R-group signature SVM models generally achieve better prediction accuracy than Free-Wilson models and that the gradient of the SVM decision function can be exploited to add Free-Wilson-like interpretability to nonlinear SVM models. The current study explores extensions of the R-group signatures by including topological distance information between the core scaffold and the signature substructure, with the purpose of improving the gradient-based R-group contributions and providing visualization of SAR within R-groups. We investigate the effect of different signature encoding schemes on the calculated R-group contributions and compare these with Free-Wilson R-group contributions. Though the main focus of the extensions presented in this study is on interpretability of the models, we also test their predictive power on compounds in external test sets, both within and beyond the Free-Wilson prediction domain.

At AstraZeneca the SignSVM-RG method has been integrated into a Web-based platform (SARPlatform) to facilitate model building, predictions, and interpretation. We use this platform to demonstrate how SVM model gradient information is applied to visualize SAR within a series of congeneric compounds.

■ METHODS

The SARPlatform. All calculations presented in this study were performed with an in-house developed Web-based platform for fragment-based SAR analysis. This application was designed to provide an easy workflow for SAR analysis on compound sets sharing a common scaffold and automate fragmentation, model building, predictions, and visualization of results. Currently, two different QSAR methods are supported; the standard Fujita-Ban/Free-Wilson method[10] and our SignSVM-RG method.[27] Figure 1 shows the workflow of the two QSAR methods. Both methods require consistent and accurate R-group decomposition data, which is provided by a novel fragmentation approach implemented within the platform using the OpenEye OEChem toolkit.[28] Stereochemical information of both core attachment atoms and R-group atoms is retained, and a consistent R-group assignment is made in cases where the core has two or more symmetrically interchangeable attachment points. The symmetry correction is a heuristic similarity-based approach where R-groups at interchangeable core attachment points are reassigned to maximize the overall LINGO[29] similarity of the fragments at each position. As a result, we place similar R-groups at the same attachment point. Such corrected assignments should minimize noise in Free-Wilson models, and the heuristic method has been proven to work well in practice. R[30] accessed through the Rpy2[31] interface was used for performing MLR and PLS analysis within the Free-Wilson method. Any linear dependencies in the Free-Wilson matrix were identified and resolved when applying MLR for model building. The SignSVM-RG method was implemented in an in-house program using the publicly available CDK[32] toolkit to generate R-group signature descriptors. LIBSVM[33] was used to build models and perform subsequent predictions using the radial basis function (RBF) kernel. Various tools are used within the platform for visualizing and analyzing model results. In this study we used the OEDepict toolkit[34] for chemical structure depiction and highlighting activity contributions of fragments within the R-groups (SigSVM-RG). Tibco Spotfire[35] is connected to SARPlatform and used as an interactive platform for model evaluation, outlier analysis, and comparisons of R-group contributions.

Signature and R-Group Gradient Contributions. The concept of using the gradient of a SVM decision function to determine contributions from R-groups in a set of congeneric compounds has already been described in our previous publication,[27] and we will briefly summarize the method here.

The SignSVM-RG method is based on the work of Carlsson et al.[36] showing that the gradient of a QSAR model can provide interpretability of nonlinear machine-learning methods. The gradient, calculated from numerical partial derivatives of the SVM decision function, represents the steepness of the modeled property space for a given attribute and can therefore be used to quantify the contribution of this attribute to the model. The SignSVM-RG method uses R-group signatures as attributes and calculates numerical gradients for nonlinear SVM models to determine the individual R-group contributions to the modeled endpoint. This is done by first calculating gradient contributions for all signatures in a training set compound and then adding up the contributions from all signatures present in each R-group separately. The occurrence number of each signature in a given R-group is used as a weighting factor in the summation. This procedure is repeated for every compound in the model training set, and the final contribution of a given R-group in the series is then calculated as an average over all compounds in the training set containing this group. In addition to the averaged R-group gradient contribution, minimum and maximum values and standard deviations are also calculated for R-groups occurring multiple times in the series. The gradient-based R-group contribution provides an easy way of interpreting the model by ranking R-groups on their respective contribution to the modeled end point. In this study we utilize the so-called atom summarized gradient contributions, which is calculated with the following equation

$$C_i = \sum_{j=1}^{k} g_j \tag{1}$$

where C_i refers to the gradient contribution of atom i which appears in k signatures. For those k signatures, g_j is the gradient value for signature j. Thus, all R-group atoms in a given compound are assigned a distinct contribution value by summing up the gradient contributions from all signatures in which each atom appears. The atom-based contributions can subsequently be projected onto the structures to facilitate interpretation of the model by highlighting specific substructures within R-groups in individual compounds, contributing either positively or negatively to the modeled end point value. Such visualization can be regarded as a more detailed view of the overall R-group gradient contribution, giving further information on which fragments within the R-groups are responsible for their relative ranking within the model. Although not discussed in this paper, the gradient calculations can also be applied to compounds in the prediction set to interpret the predictions of new R-groups at a substructure level.

Distance Encoded R-Group Signature Descriptors. The SignSVM-RG method uses fingerprints based on the signature descriptor.[26] The signature descriptor consists of a set of canonical strings describing the neighborhood of every atom in a compound. The number of atoms and bonds encoded in each signature is determined by the so-called height of the signature, where a height-N signature encodes a center atom and all neighboring atoms up to a distance of N bonds away from the center atom and their bond types. Typically the height-N signature fingerprint will include signatures of varying heights, ranging from minimal height (by default is zero) to the maximal N height, and is represented by a sparse vector where each element is an integer describing the occurrence number of each signature in the compound.

As we have described before, in the SignSVM-RG method a compound is solely described by its R-groups. Since all the structural variation in a congeneric series is in the R-groups, the common core fragment part can therefore be left out of the signature generation. Comparing to the R-group signature implementation in our previous work, the main difference is the introduction of signature distance encoding. The topological distance between the signature center and the core fragment, calculated as the number of bonds between the signature center atom and the attachment point to the core part, is utilized in the current implementation. This distance information is incorporated by adding an extra distance label in the signature string. This addition to the signature enables differentiation of identical fragments appearing at different positions around the core as well as in the same R-group, which should provide more subtle structural information for signatures than the original implementation and therefore improve the quality of the calculated R-group contribution.

A pictorial description of distance encoded R-group signatures for an example compound can be found in Figure 2, showing how distance encoded signatures can be used to distinguish the same fragment occurring multiple times within a compound. For example, fragments b, c, and d in Figure 2 would all be described by the same signature and therefore also have the same calculated gradient contribution if no distance information is included. The ability to separate signatures describing the same chemical structure but having different distances to the attachment point of the core structure should be important for the calculation of gradient-based R-group contributions, especially in the case when using smaller signatures to describe large R-groups. The influence of this

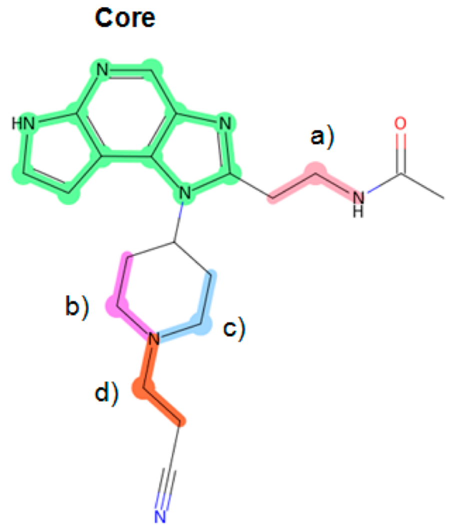


Figure 2. Illustration of distance encoded R-group signature concept for an example compound from the TYK2 series. The core scaffold is highlighted in green, and four occurrences of the same CCN fragment in the R-groups are highlighted and marked with labels a) to d) in the figure. These fragments are represented by three different height-1 distance encoded R-group signatures: [R2:C]([R2:C][R2:N])|2|(a), [R1:C]([R1:C][R1:N])|3|(b and c), and [R1:C]([R1:C][R1:N])|5|(d), where an R-group label is attached to each atom in the signature and the shortest distance (number of bonds) from the signature center atom (marked with a ball) to the core attachment atom is appended to the end of respective signature.

new encoding scheme on the predictive power and interpretability of the model is the focus of the current study.

Another addition to our previous study is that chirality is encoded into the R-group signatures, using a simple scheme where chiral carbons are replaced with either Se or Te atom labels if the stereochemistry is left- or right-handed respectively with respect to the Cahn-Ingold-Prelog (CIP) stereochemistry assignment rules in the OEChem toolkit. In order to distinguish models generated from different encoding schemes we use the following notation for a given model: sign_N-M, where N and M denote the smallest and largest height of the signatures in the descriptor respectively. A suffix "+d" in the label is used to denote that distance encoding was applied in the signature generation. The stereochemistry encoding is applied to all signature descriptors used in this study.

Matched Squares. We introduce the concept of matched squares as a simple way to check the consistency of the input data and additivity of R-group contributions. A matched pair is a set of two compounds (A→B) linked by a single-group transformation at an R-group position. A matched square is a pair of matched-pairs (A→B) and (D→C) which display the same transformation and are further linked by a second group transformation at a different position, creating the matched pairs (A→D) and (B→C). Examples of matched squares can be seen in Figure 7 (to be discussed in detail later). Looking at a property like potency, the change in pIC50 (ΔpIC50) for each of the two 'horizontal' transformations (A→B and D→C) should be the same. Likewise, the two 'vertical' transformations (B→C and A→D) would have the same values as they involve the same transformation, if the SAR is linearly additive. The absolute differences in ΔpIC50 (designated as ΔΔpIC50) for the horizontal and vertical transformations can be added up to give a metric for ranking the matched squares — a large number would indicate a discrepancy in the pIC50 changes observed for either or both of the transformations. In practice, because of measurement error the sum of the ΔΔpIC50 differences will generally not be zero even when SAR is additive and a threshold is used to single out the worst matched squares. Matched squares serve to identify compounds that exhibit nonadditive SAR and provide a quick way of identifying potential outlier compounds. Such compounds should ideally not be included for building Free-Wilson type models that rely on SAR being additive as they add noise to the model.

Data Sets and Models. In order to test the influence of different encoding schemes for the R-group signatures on models and calculated R-group gradient contributions, we have reused the eleven experimental data sets from our recent publication.[27] These data sets are from published patents extracted from the GOSTAR database[37] and contain between 230 and 1228 pIC50 or pKi values for eleven different targets. More details on the data sets can be found in Table 1. The R-group decomposition was performed using the core structures shown in Figure 3. The current R-group decomposition implementation does not allow for splitting of ring bonds, and for this reason the MMP2 series was not included in this study. For the same reason 104, 2, and 4 compounds were excluded from the IL4, MGLL, and MAPK14 series respectively, and only the remaining compounds were used in the analysis. Each data set was randomly split into training and test sets with a 4:1 ratio, and the compounds in the test set are used for external validation of the models. Some of the series contain duplicate entries of compound SMILES with different measured activity values. In these cases, only the highest activity value was included in the QSAR analysis, and the rest were discarded. Free-Wilson models were built using either MLR or PLS regression, where the latter method was used for series where many R-group coefficients could not be determined due to linear dependencies in the Free-Wilson matrix. Different encoding schemes for the R-group signatures were used to build SVM models. A SVM model using RBF kernels requires two parameters, C and gamma. For all models the local optimal C and gamma values were obtained by a grid search routine using a 5-fold cross-validation on the training set compounds to maximize the cross-validation correlation coefficient.

Table 1. Data Sets Used in This Study

data set/ target	compd description	data type	no. of compds	patent
CDK5	inhibitor	IC_{50}	230	US 20040224958 A1
IL4	inhibitor	IC_{50}	665	WO 2006/133426 A2
JAK1	inhibitor	K_i	921	WO 2011/086053
MAPK14	inhibitor	IC_{50}	610	EP 1500657 A1
PIK3CA	inhibitor	IC_{50}	304	WO 2010/139731 A1
TYK2	inhibitor	K_i	920	WO 2011/086053 A1
F7	inhibitor	IC_{50}	365	US20050043313
GNRHR	antagonist	IC_{50}	198	WO20020358
MGLL	inhibitor	IC_{50}	1228	WO2010124082
PRSS2	inhibitor	IC_{50}	339	US7119094

It is worth noting that although we have reused the same data sets from our previous publication, the models presented here cannot be directly compared with the previous ones since some models are based on different R-group decomposition data using different cores and also in some cases on reduced compounds sets for the reasons explained above. Furthermore, another difference is that in this implementation we encoded stereochemistry within the R-group's signatures (where relevant for the model).

RESULTS AND DISCUSSIONS

Signature Encoding Schemes. The main objective of this study is to analyze the effect of different signature encoding schemes within the SignSVM-RG method on the predictive power and interpretability of the models. As a secondary objective, we ascertain the stability of these nondeterministic models across repeated builds. Various signature encoding schemes were used to build QSAR models on the ten data sets to study the influence of different signature heights and encoding of distance information in the signatures on the calculated R-group gradient contributions. R-group contributions from Free-Wilson models built on the same R-group decomposition data were used as a reference for this analysis, and linear regression was applied to compare the two different R-group contribution metrics. Although we are comparing a linear and a nonlinear model, the two methods should give a similar R-group ranking within each R-group position if the biological activity in the series can be reasonably well described by additive contributions from the R-groups.

Model Stability. Due to the random sampling of compounds employed in the grid search protocol for finding optimal C and gamma parameters for SVM models, different best combinations can be obtained in separate runs on the same training set which will in turn influence the calculated gradient-

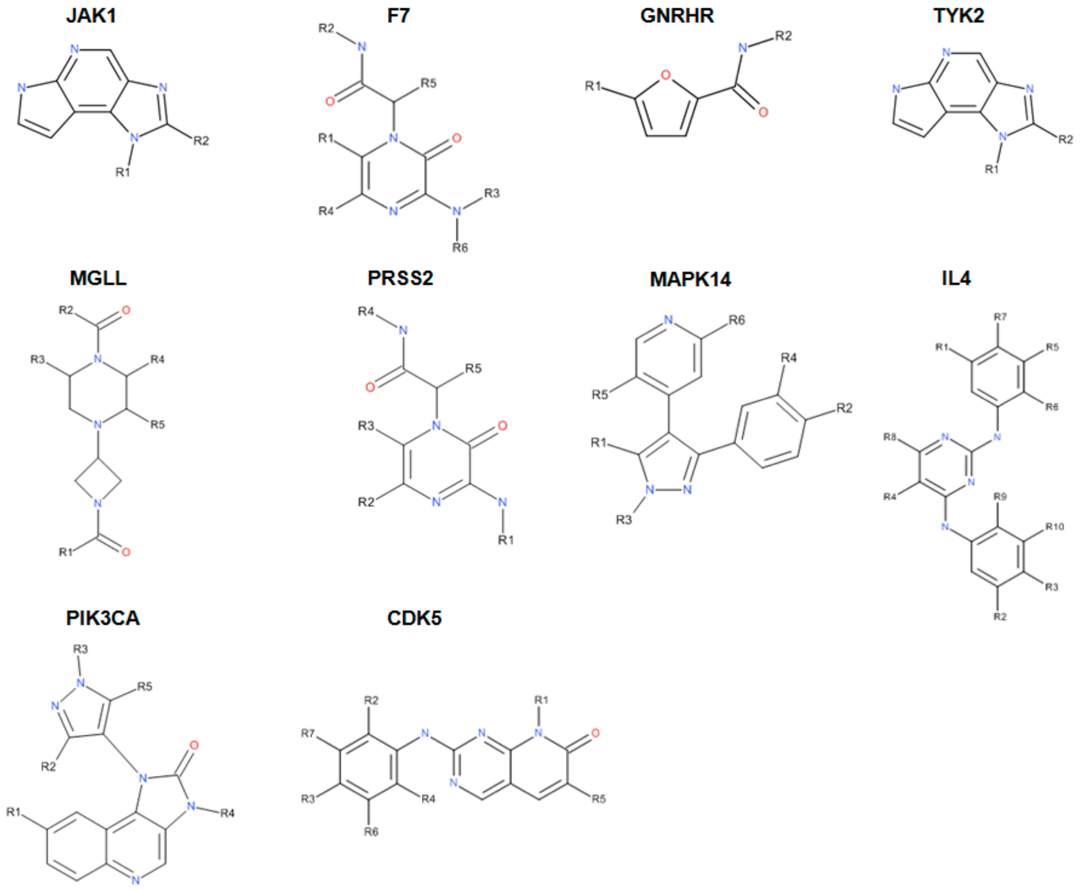


Figure 3. Scaffolds and R-groups for the ten data sets used in this study.

based R-group contributions. Therefore, ten separate SVM models were built for each signature encoding scheme and data set (300 models in total) to test the sensitivity of the gradient contributions on the SVM model parameters. Linear regression analyses between SignSVM-RG and Free-Wilson R-group contributions were performed for each SVM model, data set, and R-group position separately. Figure 4 shows average linear regression R^2 for all R-group positions containing at least ten distinct substituents in a given data set. The error bars describe the variability (standard deviation) of the calculated R^2 values over the ten SVM models, and the data points in Figure 4 are sized after the average number of heavy atoms in the R-group position, which ranges from two to fifteen atoms in these data sets. The complete data table for this analysis can be found in the Supporting Information (Table S1).

Optimal Signature Descriptor Set. In our previous study, sign_0−3 R-group descriptors were used in all analyses, so the first aspect we wanted to investigate is how signature heights influence the calculated R-group gradient contributions. The zero height signatures only contain the center atom itself (i.e., small substructures) and will in general occur more frequently than those higher height signatures (i.e., larger substructures) in R-groups and therefore have a high weight in the calculation of the R-group gradient contribution (see methods). The results in Figure 4 show that the sign_0−3 models give the overall worst correlation with Free-Wilson R-group coefficients of the three different R-group descriptors. Sign_1−3 descriptors were examined as well (results not shown), and an increased correlation with Free-Wilson coefficients over sign_0−3 models could be seen in a great majority of the data sets, especially for the larger R-groups, showing that the 0-height signatures introduce noise in the calculation of the R-group gradient contributions. Comparing the results from sign_0−3 and sign_0−3+d models (in Figure 4), in which the distance information is encoded, an increased correlation with Free-Wilson R-group coefficients can be seen for the distance encoded models in almost all data sets. The biggest improvements can be seen for the larger R-groups with more than 10 atoms (MGLL-R1, PRSS2-R4, TYK2-R1, etc.) on average, whereas only minor differences can be seen for the smaller substituents, which supports the assumption that distance encoding is important when calculating gradient-based contributions using smaller signatures to describe larger substituents. Overall, the signature encoding scheme that is in best agreement with Free-Wilson R-group contributions is sign_1−3+d, which yields correlation coefficients greater than 0.7 for all R-groups and close to perfect correlation (R^2 > 0.9) for several R-groups across different data sets.

The sensitivity analysis further shows more stable gradient contribution results of the sign_1−3+d models compared to the models based on descriptors including height-0 signatures. This suggests that the sign_1−3+d signature is the preferred alternative for interpretation of the model through gradient-based R-group contributions. We also explored the inclusion of signatures greater than height-3 in the distance encoded R-group descriptors for some of the data sets, but no further improvements were obtained in the correlation with Free-Wilson coefficients (results not shown).

Comparison to Free-Wilson. Figure 5 shows correlation plots of Free-Wilson and SignSVM-RG contributions for the

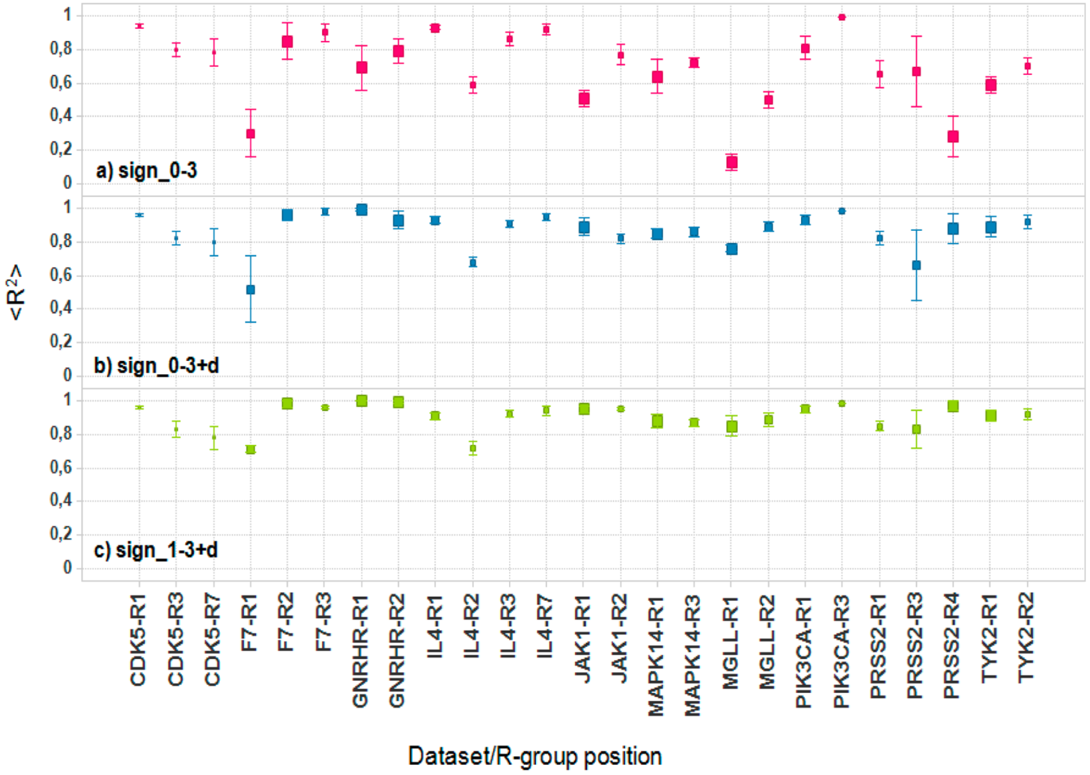


Figure 4. Correlation coefficients (R^2, Y axis) from linear regression between Free-Wilson R-group contributions and SignSVM-RG gradient contributions for the ten data sets using different signature encoding schemes. The plot shows mean value and standard deviations of R^2, corresponding to ten different SVM models with various C and gamma parameters generated by grid search, for different data set/R-group position combinations. For most of the data sets, the conventional MLR was used to build Free-Wilson models, while for IL4, MAPK14, and MGLL data sets no reasonable MLR model could be found and PLS regression using 20 components was used to build Free-Wilson models. The size of data point is proportional to the average heavy atom number at the R-group position.

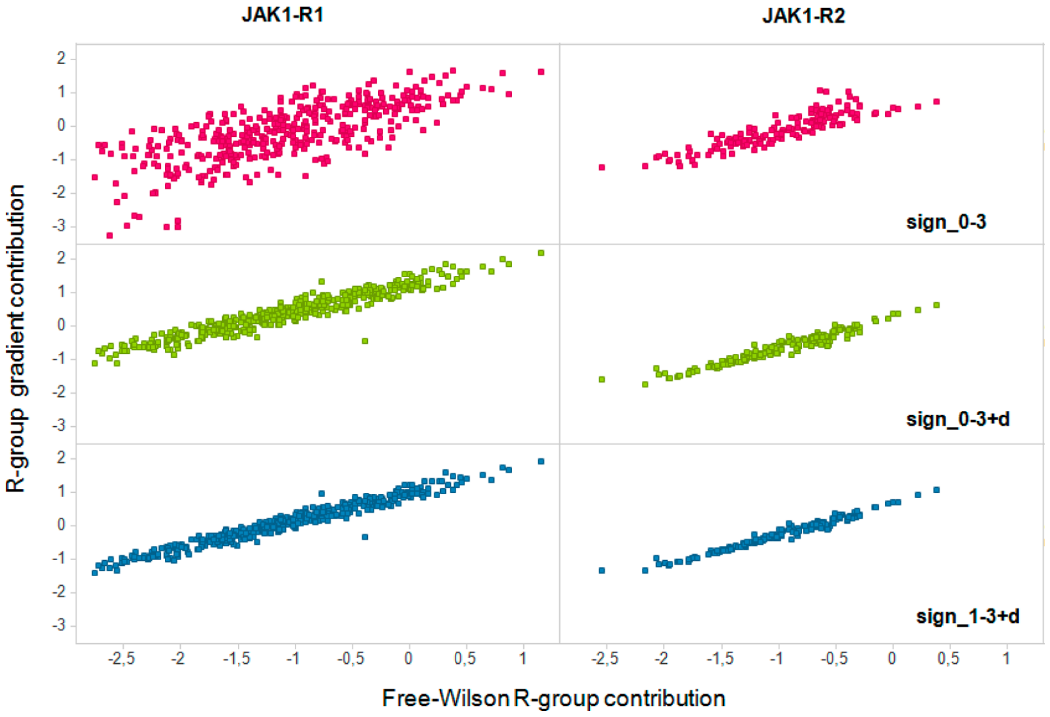


Figure 5. Free-Wilson R-group contributions vs R-group gradient contributions from SVM JAK1 models based on three different signature encoding schemes.

JAK1 data set, using SVM models based on the three previously discussed signature versions. The compounds in this data set have two larger R-groups around the core shown in Figure 3 with on average thirteen (R1) and seven (R2) atoms respectively. Figure 5 shows the significant improvements in correlation with Free-Wilson R-group contributions for the calculated R-group gradient contributions from the distance encoded SVM models. In agreement with the conclusions presented above the most drastic improvements can be seen for the larger substituents at the R1 position, with an increase in linear regression R^2 from 0.54 to 0.95 for the sign_0−3 and sign_1−3+d models respectively. These results validate the usage of SVM model gradients employing distance encoded signatures as a metric for comparing activity contributions of the substituents within a given R-group position. The ranking of R-groups occurring frequently in data sets and giving nonadditive contributions to activity is likely to be different between the two methods and may explain the lower correlation seen for some data sets in this analysis. This is discussed further in the next section.

Predictive Power. Even though the distance-encoded R-group descriptors improve the interpretability of the models, a remaining question is whether or not it can be translated into higher predictive power of the model. To test the predictive power of the models based on the three versions of descriptor discussed previously, both cross-validation and external validation was performed, and the results are presented in Tables 2−4. The results presented in these tables are based on the SVM model for each signature encoding scheme which gave overall highest prediction accuracy on the complete external test set. The complete data table for the external validation results using the 300 models from Figure 4 can be found in Table S2 in the Supporting Information.

Table 2. 10-Fold Cross-Validation Results for SignSVM-RG Models Using Different Signature Encoding Schemes

	sign_0−3		sign_0−3+d		sign_1−3+d	
data set	R^2	res. std. error	R^2	res. std. error	R^2	res. std. error
GNRHR	0.37	0.54	0.41	0.51	0.39	0.50
JAK1	0.58	0.41	0.58	0.40	0.56	0.41
F7	0.69	0.50	0.72	0.47	0.71	0.46
TYK2	0.64	0.43	0.64	0.42	0.61	0.42
IL4	0.60	0.31	0.61	0.31	0.61	0.31
MGLL	0.55	0.59	0.54	0.59	0.53	0.59
MAPK14	0.30	0.57	0.32	0.54	0.33	0.52
CDK5	0.48	0.43	0.45	0.47	0.44	0.48
PIK3CA	0.40	0.31	0.39	0.31	0.37	0.31
PRSS2	0.51	0.43	0.54	0.39	0.53	0.39

The 10-fold cross-validation results of the models in Table 2 show that the models are fairly similar. Some variations can be observed, but no clear trend can be seen in the differences between the sign_0−3, sign_0−3+d (distance encoded signatures with height from 0 to 3), and sign_1−3+d (distance encoded signatures with height from 1 to 3) models in these data sets. Table 3 shows results of various models on the subset of external compounds which are within the Free-Wilson domain (i.e., the set of compounds which is able to be predicted by a Free-Wilson model) as well as predictions from corresponding Free-Wilson models. The SVM-based predictions for the remaining compounds outside the Free-Wilson domain are presented in Table 4 (compounds in this subset have at least one R-group not present in the training set). In agreement with our previous study,[27] all three SignSVM-RG models generally perform slightly better than the Free-Wilson models on these data sets. Comparing the performance of the SVM models based on the three different signature descriptors in Table 3, some variations can be seen, but no clear trend can be drawn from these results. Although the results in Table 4 are similar for the three different signature versions in several data sets, the sign_0−3 model actually performs better when predicting compounds with novel R-groups in some data sets (GNRHR, IL4, and PIK3CA). The average R^2 for data sets in Table 4 for sign_0−3, sign_0−3+d, and sign_1−3+d models

Table 4. Predictions on Compounds in the External Test Sets Outside the Free-Wilson Domain

		SVM_sign_ 0−3		SVM_sign_ 0−3+d		SVM_sign_ 1−3+d	
data set	no. of compds	R^2	res. std. error	R^2	res. std. error	R^2	res. std. error
GNRHR	34	0.34	0.63	0.23	0.67	0.23	0.67
JAK1	119	0.47	0.48	0.46	0.48	0.49	0.47
F7	29	0.66	0.59	0.70	0.56	0.69	0.58
TYK2	113	0.54	0.46	0.54	0.45	0.53	0.46
IL4	8	0.67	0.28	0.58	0.31	0.56	0.32
MGLL	133	0.55	0.66	0.56	0.66	0.53	0.68
MAPK14	104	0.39	0.59	0.35	0.60	0.34	0.61
CDK5	2	-	-	-	-	-	-
PIK3CA	33	0.48	0.34	0.32	0.39	0.20	0.43
PRSS2	30	0.54	0.42	0.51	0.44	0.47	0.45

Table 3. Predictions on the Free-Wilson Test Sets

		Free-Wilson		SVM_sign_0−3		SVM_sign_0−3+d		SVM_sign_1−3+d	
data set	no. of compds	R^2	res. std. error	R^2	res. std. error	R^2	res. std. error	R^2	res. std. error
GNRHR	5	0.88	0.30	0.86	0.39	0.85	0.41	0.85	0.41
JAK1	66	0.81	0.37	0.79	0.33	0.81	0.32	0.77	0.35
F7	44	0.64	0.46	0.73	0.40	0.68	0.44	0.69	0.43
TYK2	71	0.73	0.40	0.80	0.38	0.79	0.39	0.79	0.39
IL4[a]	99	0.51	0.34	0.59	0.36	0.61	0.35	0.61	0.35
MGLL[a]	113	0.59	0.59	0.64	0.57	0.66	0.56	0.66	0.55
MAPK14[a]	14	0.11	1.12	0.23	0.65	0.10	0.71	0.16	0.68
CDK5	43	0.57	0.40	0.63	0.36	0.64	0.36	0.65	0.35
PIK3CA	28	0.51	0.23	0.40	0.28	0.42	0.27	0.42	0.27
PRSS2	38	0.71	0.32	0.85	0.24	0.80	0.27	0.79	0.28

[a] PLS-regression using 20 components was used to build Free-Wilson models.

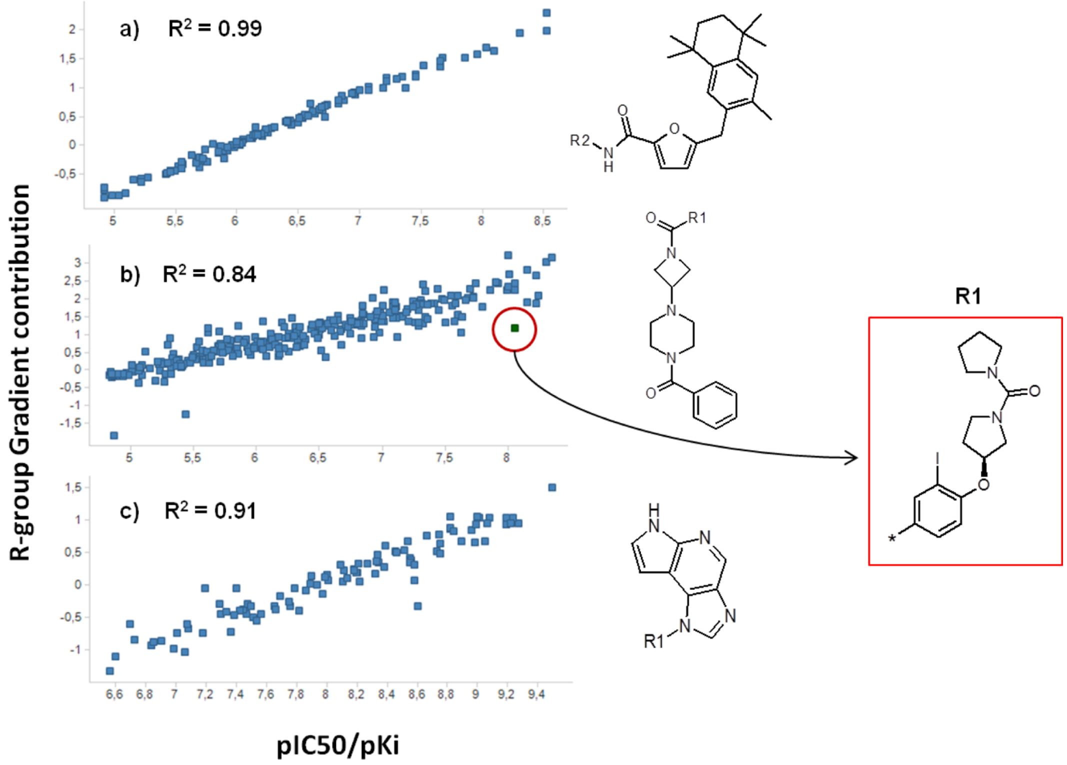


Figure 6. Calculated gradient contributions of different R-groups versus measured activity values of the molecular match pair compounds where all other substituents around the core are the same within each series. These plots show correlation for (a) 117 different R2 groups from the GNRHR series, (b) 282 R1 groups from the MGLL, and (c) 95 R1 groups from the JAK1 set. All models are built using sign_1−3+d descriptors. The common scaffolds of the compounds are shown to the right of the respective plot, and one example of an R-group discussed in the text is highlighted in the box.

are 0.52, 0.47, and 0.45, respectively. This suggests that, compared to the nondistance encoded descriptors, the distance encoded descriptors lose some information that modestly decreases the predictive power.

A possible reason for this is that the signatures that include distance encoding are less generic than nondistance-based signatures, since chemically equivalent signatures will be categorized as different descriptors according to their topological bond distance to the attachment point (illustrated in Figure 2). Some signatures having a certain distance to the attachment point may only exist in the R-groups of test set compounds but not in those of the training set, and as a result their contribution to bioactivity will be not be encapsulated in the model. With the less specific non-distance-encoded signatures, the test set compounds will likely have fewer descriptors which are not included in the training set. These results suggest that the improvement in the interpretability of the models using distance encoded signatures is at the expense of a slight decrease in their ability to extrapolate.

Additivity of R-Group Contributions. Another way of validating the method without using Free-Wilson coefficients as reference values is to use so-called molecular matched pair compounds which have different R-groups at only one substituent position and identical R-groups at all other positions. Figure 6 shows correlations between gradient contributions for the varying R-groups and the experimental pIC50 or pKi values for matched pair compounds for GNRHR, MGLL, and JAK1 models using sign_1−3+d signature descriptors. Since all other R-groups are the same within each compound set, there should be a strong correlation between the gradient contribution of the varying R-group and the measured activity of the compound when R-group contributions are additive within the series. Figure 6a shows the gradient contributions of 117 different R2 groups from the GNRHR set with the same R1 substituent plotted against the measured pIC50 values. A near to perfect correlation (R^2 = 0.99) can be seen for these compounds. More variability, but still a good overall correlation, can be seen for the MGLL and JAK1 series. Many of these R-groups occur multiple times in their respective series. It should be borne in mind that the R-group contributions in Figure 6 are calculated as averages over all occurrences for each individual R-group in the whole series, and the gradient for the same R-group in different compounds may vary. One outlier R-group from the MGLL series (in Figure 6b) serves as an example to show how the method can be used to investigate nonadditive SAR by examining the distribution of the R-group gradients from individual compounds in the data set. This R-group occurs in two compounds in the series and the calculated R-group gradient contributions are 0.08 (compound 2) and 2.28 (compound 4) respectively, giving the average contribution of 1.18 seen in Figure 6b. This is one of the largest deviations observed for an R1-group in this series and is an indication of either experimental errors or of nonadditive effects.

Further validation is shown in Figure 7, which depicts two different matched squares (see Methods) including these two compounds (compounds 2 and 4) containing the example outlier R-group. In the two horizontal transformations in the matched square, the compounds have the same R2 substitution and only differ in the R1 substituent. Similarly, in the two

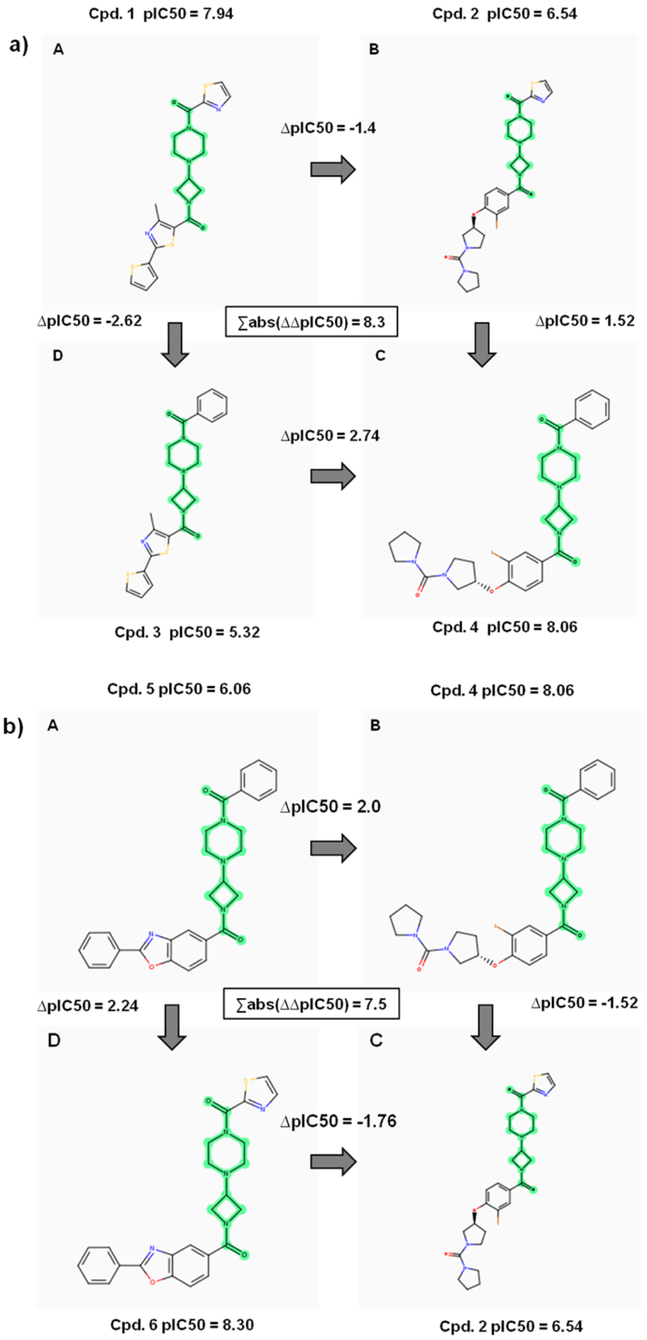


Figure 7. Two examples of matched squares including compounds 2 and 4 from the MGLL set, in which the example R-group shown in Figure 6 is contained. The core scaffold is highlighted in green, and the value in the middle of each square is the sum of the absolute ΔΔpIC50 values for the horizontal and vertical R-group transformations.

vertical transformations the R1 substitutions are the same and only the R2 substituents are different. Under the assumption that the series show perfectly additive SAR and that no experimental errors are present, the ΔΔpIC50 values for both the vertical and horizontal transformations should ideally be zero. In this case, however, a rather large difference can be seen in both cases with added absolute ΔΔpIC50 values of the horizontal and vertical R-group transformations of 8.3 and 7.5 for the two matched squares respectively. Both R2 substituents of these compounds occur frequently in the series, and the variation of the ΔpIC50 value for the "thiazole → phenyl" transformation in matched pair compounds in the series has been examined. Overall the thiazole substituent gives a much more favorable averaged contribution (1.21) compared to the phenyl group at the R2 position (0.21) in the model, while in contrast the ΔpIC50 value of the thiazole to phenyl transformation in compounds 2 and 4 is positive, showing that the R2 contributions from these two compounds are not additive within this data set. The above examples show that nonadditive R-group contributions can be captured by calculating the corresponding R-group gradient contributions in the nonlinear SVM model. These examples are the "worst" matched squares found in the MGLL series, but several other cases of clearly nonadditive R-group contributions can further be identified from the matched squares analysis on this large data set (results not shown).

Atomic Gradient Visualizations. One additional feature of the SignSVM-RG method is that the gradient information from the signatures can be visualized on an atomic level for a detailed SAR analysis to guide chemical modifications within the R-groups to optimize bioactivity. Each R-group atom can appear in many different signatures in the model, either as a center or neighboring atom. Summing up gradient contributions from all these signatures for each atom in an R-group can provide further interpretability to the complex SVM model via visualization of fragment contributions within the R-groups of each compound. Figure 8 shows visualizations of atom summarized gradient contributions for a selected set of phenyl substituted R2 groups from the sign_1−3+d GNRHR model. These depictions use a color gradient from green (positive) to red (negative) and gray in between (neutral) for the atoms to show the contribution of an atom to the activity. Another color gradient is used for the bonds based on the colors of the connected atoms. Nine molecular matched pair compounds in Figure 8 have the same R1 substituent (1,1,4,4,6,7-hexamethyltetralin), and their R2 groups are ordered by the R-group gradient contributions in a descending manner. This set of compounds enables us to relate R2 structural changes with the measured pIC50 value directly. All R2 substituents share a common phenyl scaffold which makes the interpretation of the SAR straightforward without knowing structural details of the receptor−ligand complex. Compound 8 (in Figure 8) was identified by Novartis as a potent, selective, and orally active GNRHR receptor antagonist with nanomolar affinities to human, rat, and mouse GNRH receptors after optimizing the substituents at the R2 position.[38] The R2 group for this compound is ranked as the second-best in the series by R-group gradient contributions, and it seems that the methoxy groups on the phenyl ring are important for the affinity from a comparison of the pIC50 values in the figure. This is nicely captured by the atom-summarized gradient depictions in Figure 8, in which the methoxy groups are highlighted in green, showing that they contribute favorably in all R2 groups in this set. Fragments colored in red indicate that the total contribution from the signatures at those positions is unfavorable in the model. This may lead to different interpretations in terms of chemistry. For example, in compounds 12, 13, and 15 (Figure 8) various parts of the substituted phenyl ring are colored red. This indicates that these groups provide unfavorable interactions, and other substitution patterns at these positions may result in more favorable contributions to the activity. The 2,6-dimethoxy substitution appears to be beneficial for the potency as seen for compounds 7, 8, and 9. Other ortho-para-dimethoxy

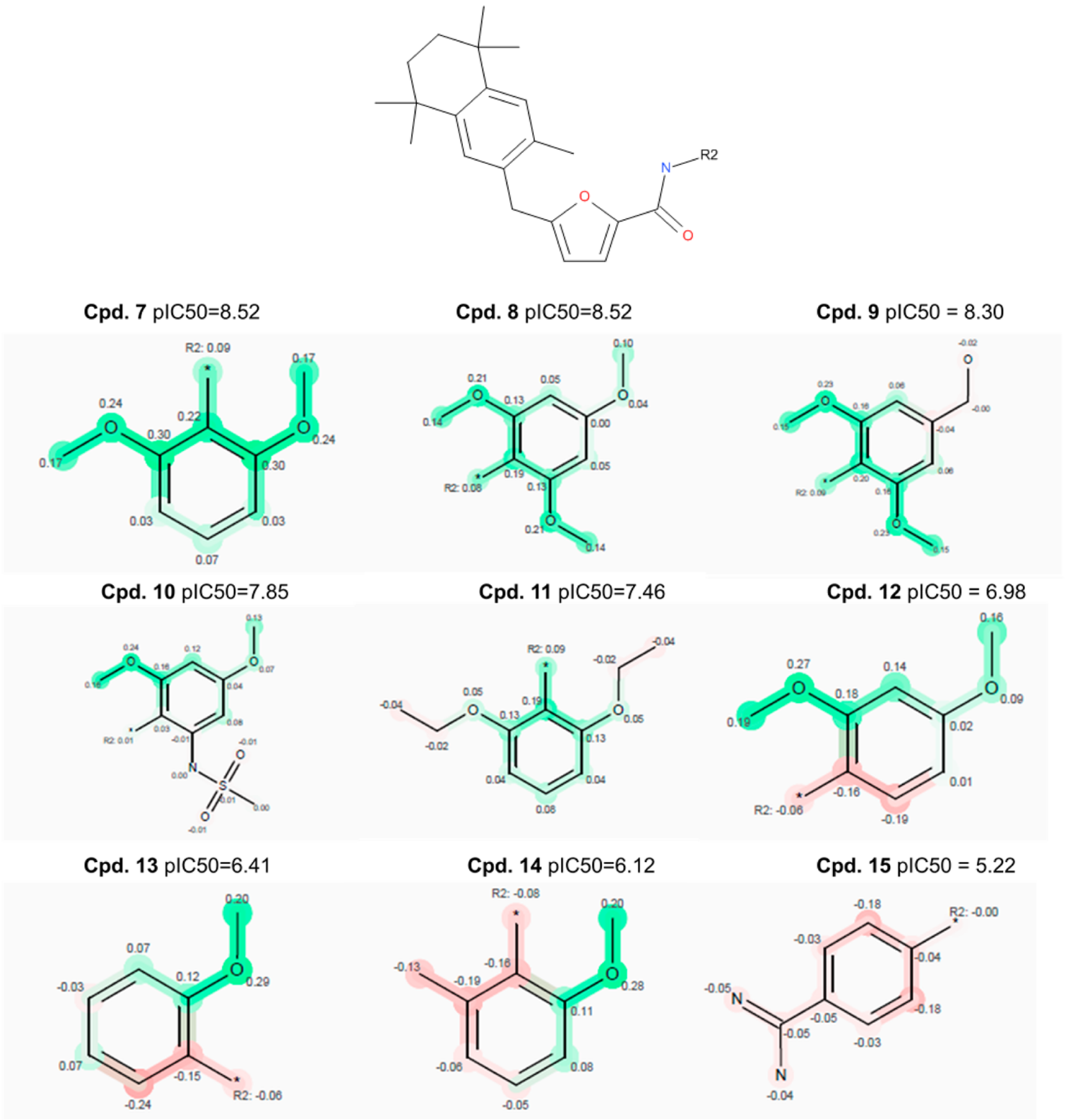


Figure 8. The atom summarized gradient contributions are demonstrated in nine selected R2 substituents from compounds in the GNRHR set. All compounds share the same R1 (1,1,4,4,6,7-hexamethyltetralin) group and the R2 groups are ordered by their R-group gradient contributions. R-group fragments are colored according to their contribution to compound potency using a color gradient from green (positive) to gray (neutral) and to red (negative). The atom summarized gradient contribution for each atom is shown in numbers, and core attachment points are marked by a star in each picture.

substituent patterns (as seen in compound 12) or mono-methoxy substitution (compound 13) are suggested to be less favorable. The 2,6-diethoxy substitution in compound 11 is detrimental to the activity as compared to compound 7, and probably the added methyl groups cause less favorable interactions of the group with the target protein and results in a decrease in binding affinity of approximately one log unit. This is illustrated in Figure 8 where the two terminal carbon atoms in both ethoxy groups are colored red and the remaining atoms of the group are less positive (less green) than the corresponding atoms of compound 7. The ortho-methyl group in the 2-methoxy-6-methyl-phenyl group in compound 14 is indicated in red, according to the model, suggesting that the major drop in potency compared to compound 7 is associated with this group. Similarly the modest drop in activity from compound 9 to compound 10 appears to be primarily due to the change of para-hydroxymethyl (9) to ortho-methanesulfonamido (10) groups. The para-amidine group in compound 15 affects the interaction of the whole benzamidine group, and the whole group is red, in agreement with the low compound potency (pIC50 5.12). It should be pointed out that the visualizations will mainly be useful for a series where the R-groups share at least some structural similarity, otherwise the R-groups will be colored uniformly and provide no extra information over the R-group ranking shown in the previous sections.

The SAR visualizations exemplified in this section show the strength of the SignSVM-RG method as a tool for SAR analysis, providing detailed interpretation of the model in a manner which should be clear and intuitive for drug designers. A further use of these visualizations is as a quick tool to investigate additivity in the series, since a separate depiction for identical R-groups in different training set compounds will be generated. Similar gradient calculations and visualizations can be applied to full compound signature SVM models,[36] but much of the detailed information given by R-group-specific signature descriptors will be lost.

CONCLUSIONS

In this paper we describe the use of nonlinear R-group QSAR based on the signature descriptor for SAR analysis in a congeneric series. The method employs a hybrid Free-Wilson/SVM approach, and we demonstrate the applicability of SVM-derived gradient contributions of signatures in visualizations for detailed SAR analysis.

As an extension of our previous method, using R-group signatures encoding the distance-to-attachment point information largely improves the accuracy of SVM model gradients and hence the interpretability of the model itself. Visualization of such gradients provides SAR information on regions within R-groups and as such offers insights that are complementary to a traditional Free-Wilson analysis. Our results show that the hybrid Free-Wilson/SVM approach with the addition of distance information in signatures remarkably improves the interpretability of nonlinear models; however, it does not necessarily improve predictive power. A possible reason for the latter is that fragments in new R-groups (i.e., not present in the training set compounds) are more likely to be missing from models using distance encoded signatures (which give rise to higher numbers of different descriptors than the nondistance encoded signatures). As a result, models using distance encoded signatures may have a more limited domain of applicability and hence a reduced ability to extrapolate compared with nondistance-encoded signatures.

The SignSVM-RG approach could be an attractive enhancement to Free-Wilson analysis and represents a powerful QSAR tool for medicinal chemists. The development of the method is ongoing, and several other extensions and additions can be envisioned. For example, the accuracy of SVM model gradients could be ameliorated by using analytical gradients of the SVM decision function, instead of the discrete numerical gradients used in the current implementation. The method can also be used to suggest new compounds in a subsequent design-make-test-analyze cycle by applying inverse QSAR based on the signatures in the model. The signature descriptor has already successfully been applied to inverse QSAR studies[39−41] and could be used to propose new substituents at a given attachment point around a scaffold.

ASSOCIATED CONTENT

Supporting Information

The detailed linear regression results for R-group contributions between different SignSVM-RG models and Free-Wilson models and comparison of model performance on full test sets in all ten runs for each data set. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Authors

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

Notes

The authors declare no competing financial interest.

REFERENCES

(1) Dickinson, M.; Gagnon, J. P. The cost of new drug discovery and development. *Discovery Med.* **2004**, *4*, 172−9.

(2) Cumming, J. G.; Davis, A. M.; Muresan, S.; Haeberlein, M.; Chen, H. Chemical predicative modelling to improve compound quality. *Nat. Rev. Drug Discovery* **2013**, *12*, 948−962.

(3) Senese, C. L.; Duca, J.; Pan, D.; Hopfinger, A. J.; Tseng, Y. J. 4D-fingerprints, universal QSAR and QSPR descriptors. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1526−39.

(4) Wold, S.; Ruhe, A.; Wold, H.; Dunn, W. J., III The colinearity problem in linear regression. The Partial Least Squares (PLS) approach to generalized inverses. *SIAM J. Sci. Stat. Comput.* **1984**, *5*, 735−43.

(5) Svetnik, V.; Wang, T.; Tong, C.; Liaw, A.; Sheridan, R. P.; Song, Q. Boosting: An ensemble learning tool for compound classification and QSAR modeling. *J. Chem. Inf. Model.* **2005**, *45*, 786−99.

(6) Mulgrew, B. Applying radial basis functions. *IEEE Signal Process.* **1996**, *13*, 50−65.

(7) Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273−97.

(8) Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5−32.

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

(10) Fujita, T.; Ban, T. Structure-activity study of phenlamines as substrates of biosynthetic enzymes of sympathetic transmitters. *J. Med. Chem.* **1971**, *14*, 148−52.

(11) 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.

(12) 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.

(13) 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.

(14) 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.

(15) 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.

(16) 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 inhibitors. *J. Med. Chem.* **2010**, *53*, 4399−411.

(17) 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.

(18) 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.

(19) 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.

(20) An, Y.; Sherman, W.; Dixon, S. Kernel-based partial least squares: Application to fingerprint-based QSAR with model visualisation. *J. Chem. Inf. Model.* **2013**, *53*, 2312−21.

(21) Myint, K. Z.; Wang, L.; Tong, Q.; Xie, X. Q. Molecular fingerprint-based artificial neural networks QSAR for ligand biological activity predictions. *Mol. Pharm.* **2012**, *9*, 2912−23.

(22) Daylight Manual; http://www.daylight.com/dayhtml/doc/theory/theory.finger.html (accessed 2013).

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

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

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

(26) 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.

(27) Chen, H.; Carlsson, L.; Eriksson, M.; Varkonyi, P.; Norinder, U.; Nilsson, I. Beyond the scope of Free-Wilson analysis: Building interpretable QSAR models with machine learning algorithms. *J. Chem. Inf. Model.* **2013**, *53*, 1324−36.

(28) OpenEye Scientific Software; http://www.eyesopen.com/oechem-tk (accessed 2013).

(29) Grant, J. A.; Haigh, J. A.; Pickup, B. T.; Nicholls, A.; Sayle, R. A. Lingos, finite state machinies, and fast similarity searching. *J. Chem. Inf. Model.* **2006**, *46*, 1912−8.

(30) R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing: Vienna, Austria. ISBN 3-900051-07-0; http://www.R-project.org (accessed 2013).

(31) Rpy: A simple and effective access to R from Python; http://rpy.sourceforge.net/rpy2.html (accessed 2013).

(32) Steinbeck, C.; Han, Y.; Kuhn, C.; Horlacher, O; Luttmann, E.; Willighagen, E. The Chemistry Development Kit (CDK): An open-source Java library for chemo- and bioinformatics. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 493−500.

(33) Chang, C.-C.; Lin, C.-J. LIBSVM: A library for support vector machines. *ACM Trans. Intelligent Syst. Technol.* **2011**, *2*, 1−27.

(34) OpenEye Scientific Software; http://www.eyesopen.com/oedepict-tk (accessed 2013).

(35) TIBCO Spotfire 3.1; http://spotfire.tibco.com (accessed 2013).

(36) Carlsson, L.; Ahlberg Helgee, E.; Boyer, S. Interpretation of nonlinear QSAR models applied to Ames mutagenicity data. *J. Chem. Inf. Model.* **2009**, *49*, 2551−8.

(37) GOSTAR databases 2012; GVK Bioscieces Private Ltd.: Hyderabad, India.

(38) Anderes, K. L.; Luthin, D. R.; Castillo, R.; Kraynov, E. A.; Castro, M.; Nared-Hood, K.; Gregory, M. L.; Pathak, V. P.; Christie, L. C.; Paderes, G.; Vazir, H.; Ye, Q.; Anderson, M. B.; May, J. M. Biological characterization of a novel, orally active small molecule Gonadotropin-Releasing Hormone (GnRH) antagonist using castrated and intact rats. *J. Pharamacol. Exp. Ther.* **2003**, *305*, 688−95.

(39) Churchwell, C. J.; Rintoul, M. D.; Martin, S.; Visco, D. P.; Kotu, A.; Larson, R. S.; Sillerud, L. O.; Brown, D. C.; Faulon, J.-L. The signature descriptor 3. Inverse-quantitative structure-activity relationship. *J. Mol. Graphics Modell.* **2004**, *22*, 263−73.

(40) Martin, S. Lattice enumeration for inverse molecular design using the signature descriptor. *J. Chem. Inf. Model.* **2012**, *52*, 1787−97.

(41) Ahlberg Helgee, E.; Carlsson, L.; Boyer, S. A method for automated molecular optimization applied to Ames mutagenicity data. *J. Chem. Inf. Model.* **2009**, *49*, 2559−63.