Predicting Kinase Selectivity Profiles Using Free-Wilson QSAR Analysis

Simone Sciabola,*[,†,‡] Robert V. Stanton,[‡] Sarah Wittkopp,[‡] Scott Wildman,[‡] Deborah Moshinsky,[‡] Shobha Potluri,[‡] and Hualin Xi[‡]

Laboratorio di Chemiometria, Universitá di Perugia, Via Elce di Sotto, 10, 1-06123, Perugia, Italy, and Pfizer Research Technology Center, Cambridge, Massachusetts 02139

Received April 17, 2008

Kinases are involved in a variety of diseases such as cancer, diabetes, and arthritis. In recent years, many kinase small molecule inhibitors have been developed as potential disease treatments. Despite the recent advances, selectivity remains one of the most challenging aspects in kinase inhibitor design. To interrogate kinase selectivity, a panel of 45 kinase assays has been developed in-house at Pfizer. Here we present an application of *in silico* quantitative structure activity relationship (QSAR) models to extract rules from this experimental screening data and make reliable selectivity profile predictions for all compounds enumerated from virtual libraries. We also propose the construction of R-group selectivity profiles by deriving their activity contribution against each kinase using QSAR models. Such selectivity profiles can be used to provide better understanding of subtle structure selectivity relationships during kinase inhibitor design.

INTRODUCTION

Over 500 human genes encode protein kinases, and these can be grouped into a number of subsets based primarily on sequence and structural similarities.[1,2] In general, protein kinases catalyze the transfer of the terminal phosphoryl group of ATP to specific hydroxyl groups of serine, threonine, or tyrosine residues of their protein substrates. Thus, protein kinases can broadly be considered serine/threonine kinases or tyrosine kinases or in some instances dual-specificity kinases when they phosphorylate serine/threonine as well as tyrosine residues.

Because protein kinases have profound effects on a cell, their activity is highly regulated by the binding of activator/inhibitor proteins or small molecules or by controlling their location in the cell relative to their substrates. Intracellular phosphorylation by protein kinases, triggered in response to extracellular signals, provides a mechanism for the cell to switch on or off many diverse processes.[3] These processes include metabolic pathways, kinase cascade activation, membrane transport, gene transcription, and motor mechanisms. Deregulated kinase activity is a frequent cause of disease, particularly cancer, since kinases regulate many aspects that control cell growth, movement, and death. Drugs which inhibit specific kinases are being developed to treat many diseases, and several are currently in clinical use, including Gleevec[4] (Imatinib) for chronic myeloid leukemia (CML) and Sutent[5] (Sunitinib) a multitargeted receptor tyrosine kinase for the treatment of renal cell carcinoma (RCC) as well as imatinab-resistant gastrointestinal stromal tumor (GIST) and Iressa[6] (Gefitinib) and Erlotinib[7] (Tarceva) for nonsmall cell lung cancer (NSCLC), providing proof-of-principle that small molecule kinase inhibitors can be effective drugs.

There are currently over 30 known inhibitors in clinical trials or approved for use in humans, few of which are highly specific for their kinase target with respect to the rest of the kinases. The particular set of kinases inhibited by a compound may drastically affect its therapeutic usefulness. Moreover, the majority of these molecules specificity has been determined against only a relatively small set of kinases. Previously reported studies have shown how molecular specificity varies widely among known inhibitors,[8] and this variation is not dictated by the general chemical scaffold of an inhibitor (i.e., EGFR inhibitors, belonging to the quinazoline/quinoline class, range from highly specific to quite promiscuous) or by the primary, intended kinase target toward which the particular inhibitor was initially optimized (i.e., compounds considered TK inhibitors also bind to Ser-Thr kinases and vice versa).[8]

Recent improvements in robotics, data processing, liquid handling, and detectors allow efficient screening of thousands of compounds against a panel of protein targets. The data resulting from such efforts are well suited for use with *in silico* modeling approaches. Here we use these techniques to probe kinase specificity by analyzing the propensity of a diverse set of protein kinases to be inhibited by small molecule inhibitors.

Pfizer has developed an internal Kinase Selectivity Screening platform[9] (KSS) to provide high quality selectivity data against a diverse range of kinases, thereby guiding project Structure Activity Relationship (SAR) and identifying potential safety liabilities in chemical series. Kinases in the panel were carefully selected based on bioinformatic and structural data to provide representative coverage across subfamilies within the kinome. Screening reagents and conditions are quality controlled to provide a high degree of reproducibility.

When this highly valuable source of data is properly integrated with computer-assisted methods, it can simplify the process of lead-discovery by allowing property-based

* Corresponding author phone: (617)551-3327; fax: (617)551-3117; e-mail: simone.sciabola@pfizer.com. Corresponding author address: Research Technology Center, Cambridge, MA 02139.

[†] Universita′ di Perugia.

[‡] Pfizer Research Technology Center.

10.1021/ci800138n CCC: $40.75 © 2008 American Chemical Society
Published on Web 08/22/2008

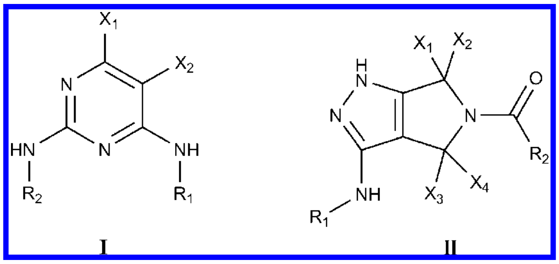


Figure 1. 2D depiction for core-structures of diaminopyrimidine and pyrrolopyrazole series. X_1, X_2, X_3, and X_4 symbolize substitutions around the core whose structure remains constant among all the compounds in the series. R1 and R2 are attachment points for R-group substitutions.

design within a series of homologues. Here, modeling can contribute to optimize pharmacological or pharmacokinetic properties via Quantitative Structure−Activity Relationship (QSAR) models. However, the success of such studies depends on the choice of an appropriate molecular characterization, through the use of informative descriptors. Usually, QSARs are derived by correlating information about activity or binding data and ligand properties, which can be one- (i.e., molecular weight, logP, properties count, or structural descriptors), two- (i.e., structural keys, hashed fingerprints),[10,11] or three-dimensional descriptors (i.e., GRID-based,[12−14] CoMFA,[15] pharmacophore fingerprints[16−18]). This can be done using any of several methods including multiple linear regression (MLR),[19] partial least-squares (PLS),[19] neural networks,[20] random forest,[21] support vector machine,[22,23] or similarity indices.[24]

In the present study, MLR with the Fujita-Ban[25−28] modification of Free-Wilson[29−33] analysis was used to derive a quantitative understanding of the selectivity data for a set of diaminopyrimidine and pyrrolopyrazole-based compounds against the protein kinases of the KSS platform. Within these two series of analogues, reliable estimations of the activity contribution for each combination of R-group/kinase were observed. Library enumeration based on the computed R-group contributions was performed in an attempt to predict more selective compounds in the virtual space of the existing monomers set, and a strong correlation of experimental versus predicted inhibition values was found for a subset of these compounds subsequently tested in the kinase selectivity panel. Finally, public protein kinase crystal structures allowed us to further validate our QSAR models by combining the information from the Free-Wilson approach with the three-dimensional (3D) structural knowledge of the target, providing more insight for kinase selectivity.

COMPUTATIONAL DETAILS

A. Training Set. The KSS is the Pfizer internal Kinase Selectivity Screening panel consisting of 45 different protein kinase bioassays selected based on bioinformatics and structural data to provide maximal coverage across subfamilies within the kinome. Thousands of compounds have been tested for IC_{50} and percent inhibition against the KSS panel in its current form, and this number is rapidly increasing. A subset of the experimental data available in the KSS was used for this Free-Wilson case study consisting of 700 compounds belonging to two different chemotypes: the diaminopyrimidine series (Figure 1, core I) and the pyrrolopyrazole series (Figure 1, core II), with 388 (R1=77, R2=183, R3=8) and 312 (R1=124, R2=87) available structures respectively, in an incomplete combinatorial matrix. These two series were chosen based on their availability as core structures in the public domain and the high number of compounds tested against the KSS panel within each series (Figure 3). While the entire KSS panel has typically been run against each compound, there remains some variability in the number of kinases profiled for each compound, depending on when the compound was profiled and when the assay was added to the panel. In order to build reliable QSAR models for every protein kinase present in the KSS, our criteria for series selection had to take into account which series consistently had a high number of compounds per kinase assay (Figure 3).

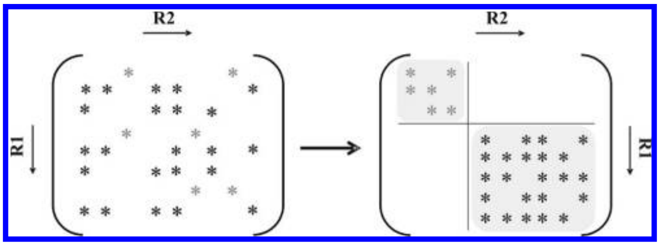


Figure 2. Structural matrix transformation. Black and gray stars represent subsets of compounds whose R-groups never cross over with each other. This allows R-group combinations to be rearranged into independent blocks and statistical analysis to be run separately within each block.

Moreover, the majority of compounds in the KSS panel were tested for percentage inhibition at two compound concentrations (1 μM and 10 μM) but were not tested for IC_{50} directly. Typical QSAR models built on IC50 data are considered more reliable and reproducible. However, due to the cost of panel screens (IC50 is normally determined from dose response experiments using from 6 to 12 doses), it is a common practice for scientists to first screen compounds in only one or two doses. Therefore, only a small fraction (~10%) of the compounds in the KSS panel was subject to final IC50 determination. In order to maximize the use of the experimental data for building QSAR models, we designed a formula to combine the percentage inhibitions at two doses (1 μM and 10 μM) into a single pIC50 value. Ekins et al. have found in their study that the IC_{50} value could reliably be predicted by a single value of percent inhibition.[34] We validated this conversion with a set of 360 compounds, representing all the compounds in the KSS panel for which both percentage inhibition and IC_{50} data were present, and we found a very strong correlation between the IC_{50} estimated from percentage inhibition and the experimentally determined IC_{50} (Figure 4). This result is indicative of the high quality of the KSS assays and also gives us confidence to use the percentage inhibition data for our QSAR study.

Data transformation was first done by applying eq I to convert percent inhibition values into IC_{50}:[34]

$$IC_{50} = C \times \frac{100 - (\text{percent inhibition at } C)}{\text{percent inhibition at } C} \qquad \text{(i)}$$

As all compounds in the KSS were tested at both 1 μM and 10 μM concentrations (C), these data were transformed

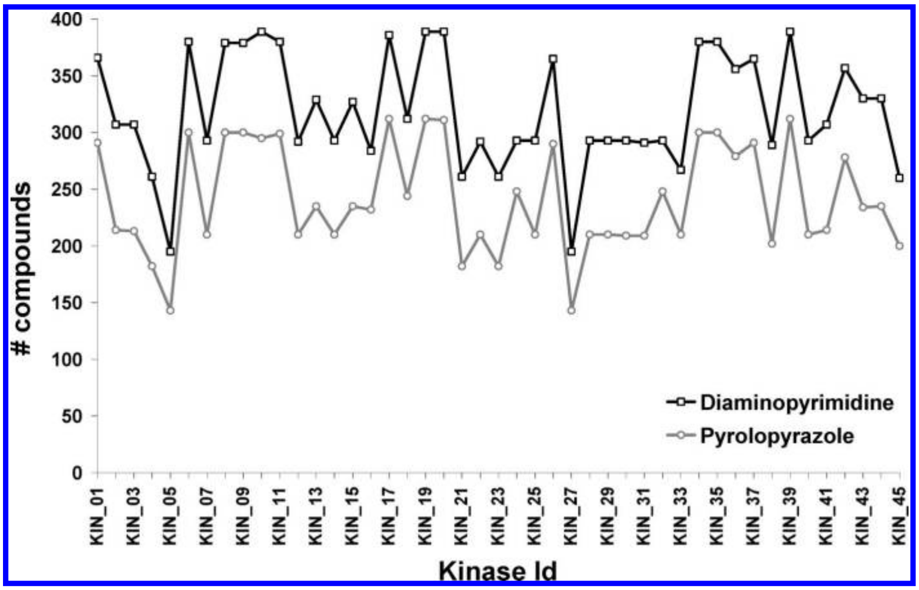


Figure 3. Profile plot showing the number of compounds tested experimentally against each kinase on the KSS panel.

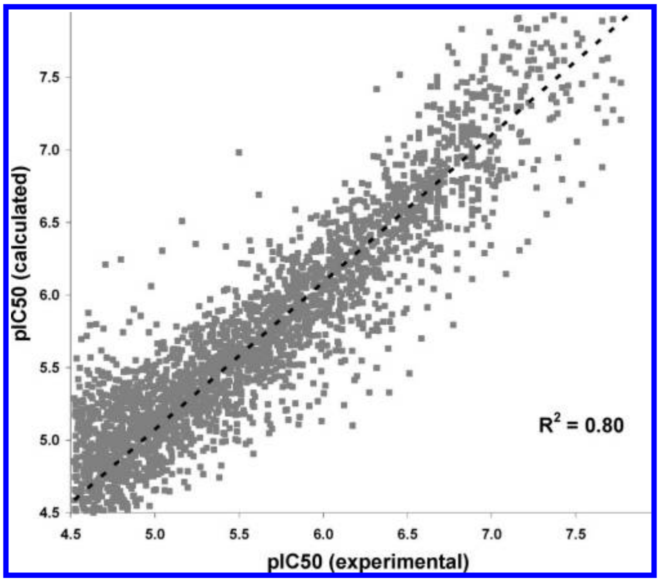


Figure 4. Activity correlation between estimated pIC_{50} from percentage inhibition (pIC_{50} calculated) and experimentally determined pIC_{50} (pIC_{50} experimental). Overall, high correlation is observed within the range of experiments.

into logarithmic scale separately and then merged by compound according to

$$pIC_{50}^{Calc} = \begin{cases} pIC_{50}@1\ \mu M, & \text{Inhib}@10\ \mu M > 99\% \\ pIC_{50}@10\ \mu M, & \text{Inhib}@1\ \mu M < 5\% \\ \dfrac{pIC_{50}@1\ \mu M + pIC_{50}@10\ \mu M}{2}, & 5\% \le \text{Inhib}@1\ \mu M \le 99\% \end{cases} \qquad \text{(ii)}$$

We decided to adopt the reported block function because in the lower range of inhibition, below 5%, there was a stronger correlation between pIC_{50}^{Calc} computed at 10 μM concentration versus experimental pIC_{50}^{Exp}, when compared to 1 μM. An opposite trend was present in the upper range of inhibition (above 99%), where pIC_{50}^{Calc} computed at 1 μM concentration tended to correlate better with experiment than that at 10 μM. For inhibition values between the previously defined cut-offs, we used the average pIC_{50}.

B. Assay Conditions. All of the KSS assays are performed in a 384-well format using either a radioactive[35,36] or Caliper[37,38] protocol. In all assays, 5 μL of 5x concentration compound in 3.75% DMSO is added to the plates. Ten μL of 2.5x enzyme in 1.25x kinase buffer (optimized for each individual kinase) is then added, followed by a 15 min preincubation at room temperature. Ten μL of a 2.5x mixture of peptide substrate (optimized for each individual kinase) and ATP in 1.25x kinase buffer are then added to initiate the reaction. Each assay is run at the experimentally determined Michaelis−Menten constant (K_m) concentration of ATP for the relevant kinase with an incubation time that was determined to be within the linear reaction time. Reactions are stopped by the addition of EDTA to a final concentration of 20 mM. Detection of phosphorylated substrate is achieved using either a radioactive method or a nonradioactive mobility shift assay format (Caliper). In the radioactive assay, tracer amounts of γ-^{33}P labeled ATP are included in the reaction, and biotinylated peptide substrates is used. After the reactions are stopped, 25 μL are transferred to streptavidin coated Flashplates (PerkinElmer). Plates are washed with 50 mM Hepes and soaked for 1 h with 500 μM unlabeled ATP before reading in a TopCount. Alternatively, for the mobility shift assay, reactions are stopped within the assay plates followed by detection of fluorescently labeled substrates on a Caliper LC3000 using a 12-sipper chip and conditions that were optimized for each kinase.

C. Free-Wilson Theory. The Free-Wilson approach was the first mathematical technique to be developed for the quantitative prediction of the Structure−Activity Relationships for a series of chemical analogs.[29] The basic idea behind this methodology is that the biological activity of a molecule can be described as the sum of the activity contributions of specific substructures (parent fragment and the corresponding substituents). It does not require any substituent parameters or descriptors to be defined; only the activity is needed. However, it must satisfy some assumptions that are not necessarily valid for all of the data set. First, it is assumed that each substituent on the parent structure makes a constant contribution to the activity, regardless of the structural variation in the rest of the molecule. Second, it is assumed that these contributions are additive and third that

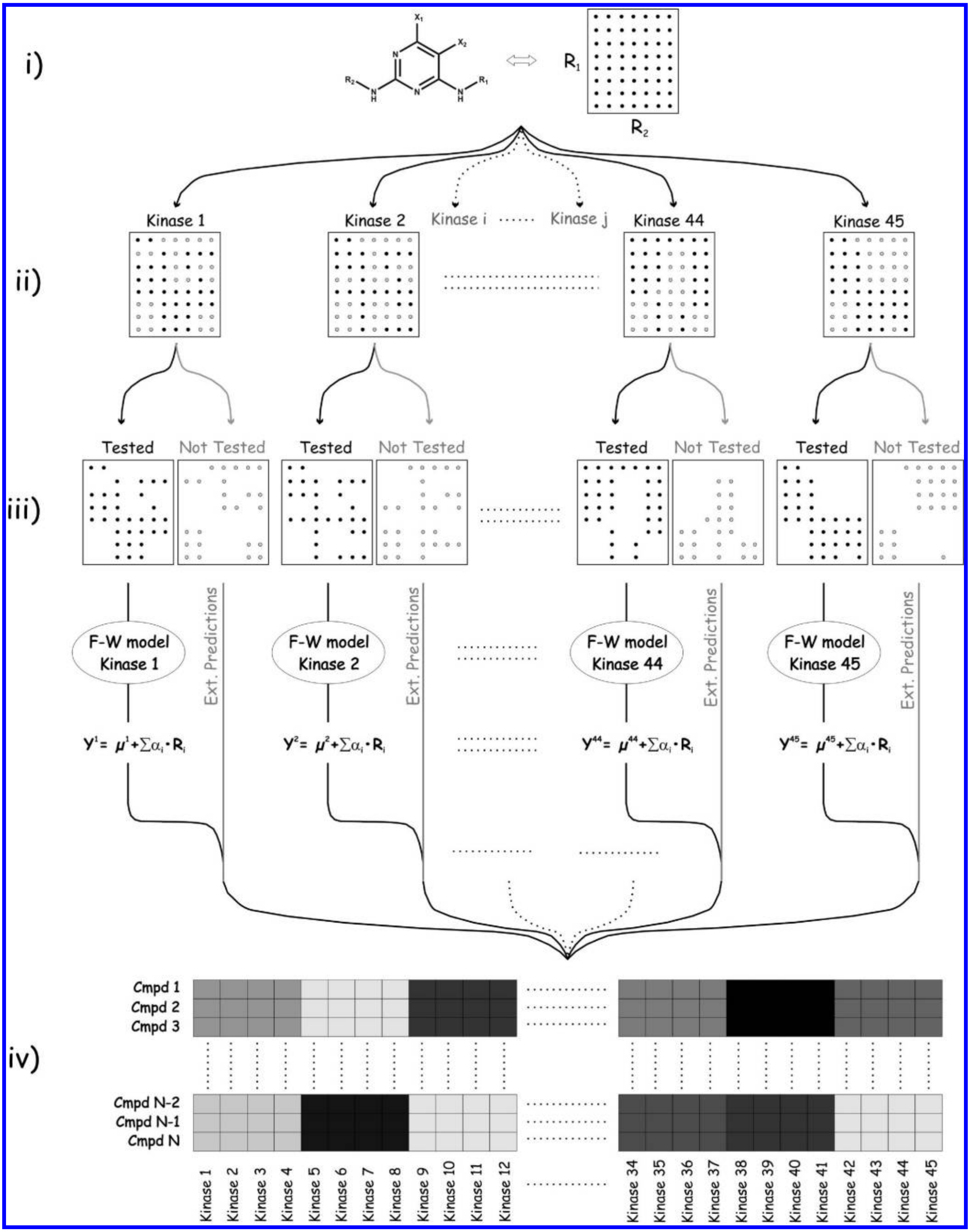
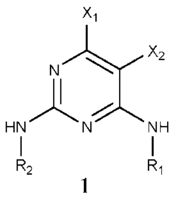
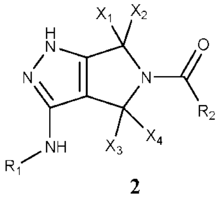


Figure 5. Free-Wilson protocol for R-group selectivity profiles. 1) In the diaminopyrimidine series, R1xR2 combinations are possible and correspond to the complete combinatorial space as defined by the R-groups in the data set. 2) However, not all the combinations were explored experimentally in the KSS panel; also different R-group combinations were tested in different kinase assays. Black circles indicate compounds in KSS, while gray circles are for compounds not yet tested. 3) The original data are split into two different sets, “Tested” and “Not Tested”. Tested compounds were used as the learning set for training each individual Free-Wilson model, giving a set of R-group activity contributions (MLR coefficients) for each different kinase model. 4) R-groups activity contributions were eventually applied to predict the kinase activity profiles for the “Not Tested” set.

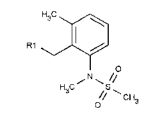
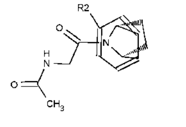
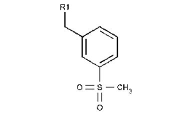
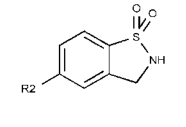
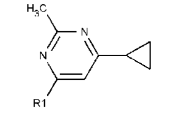
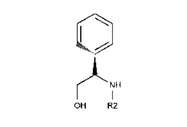
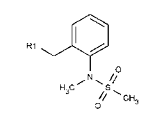
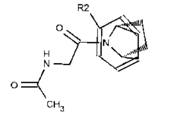
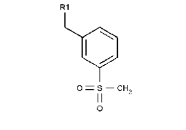
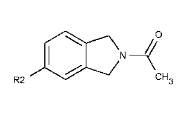
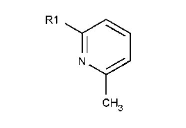
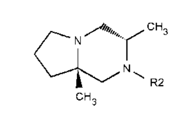
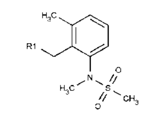
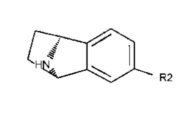
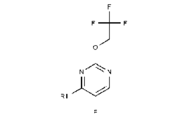
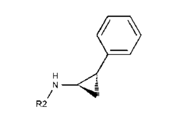
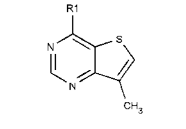
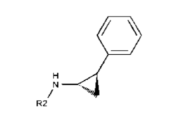
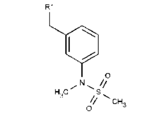
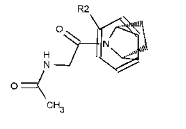
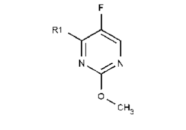
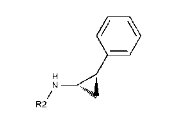
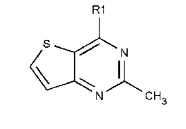
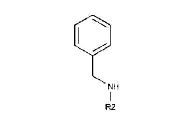
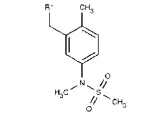
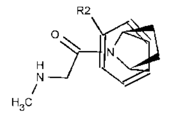
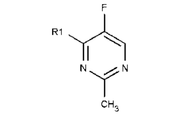
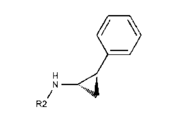
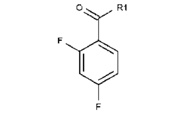
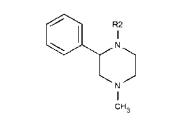
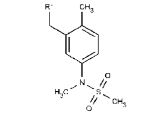
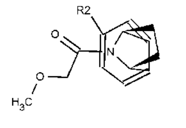
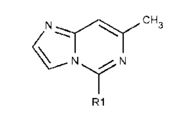
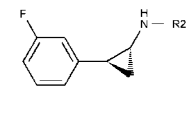
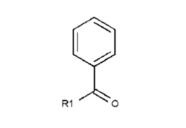
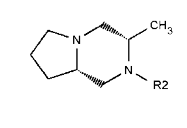
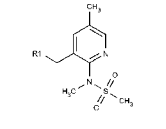
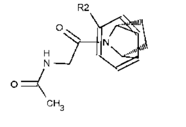
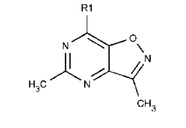
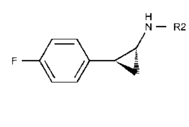
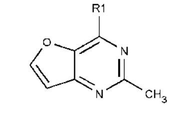
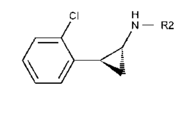
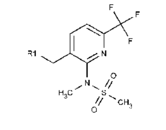
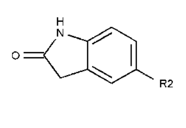
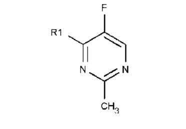
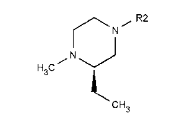
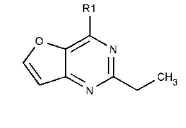
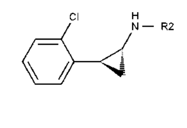
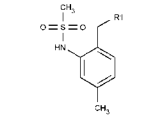
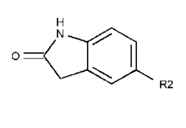
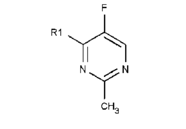
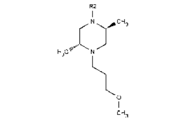
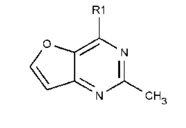
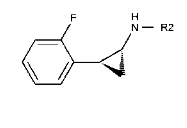
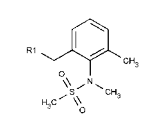
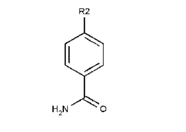
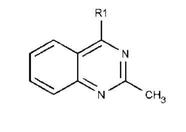
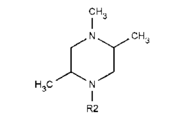
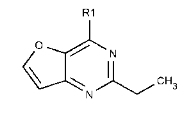
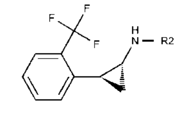
Table 1. 2D Depiction for the 30 Test Set Compounds[a]



1



2

ID	R1	R2	ID	R1	R2	ID	R1	R2
1a			1l			2i		
1b			1m			2k		
1c			2a			2l		
1d			2b			2m		
1e			2c			2n		
1f			2d			2o		
1g			2e			2p		
1h			2f			2q		
1i			2g			2r		
1k			2h			2s		

[a] X_1, X_2, X_3, and X_4 represent not changing chemical matter whose structures cannot be disclosed.

there are no interactions between the parent fragment and its substituents or between the substituents themselves. Lastly, Free-Wilson methodology can only explore the chemical space defined by the R-group combinations present

Table 2. Definition of the Most Important VolSurf Descriptors To Explain R-Group Kinase Selectivity

type	name	description
size and shape	V	Volume of the water molecule interaction field at 0.2 kcal/mol energy level.
	S	Surface of the water interaction field at 0.2 kcal/mol energy level.
	G	Globularity: ratio between the surface (S) and the surface of a sphere with the same volume (V).
	rugosity	Measure of molecular wrinkled surface; it represents the ratio of volume/surface. The smaller the ratio, the larger the rugosity.
hydrophilic regions	W1−W8	Volumes of the water molecule interaction fields at eight different energy levels: −0.2, −0.5, −1.0, −2.0, −3.0, −4.0, −5.0, and −6.0 kcal/mol.
	W1N−W8N	Volumes of the amide molecule interaction fields at eight different energy levels: −0.2, −0.5, −1.0, −2.0, −3.0, −4.0, −5.0, and −6.0 kcal/mol.
	Iw1-Iw8	Integy moments: distances between the center of mass of the molecule and the center of the hydrophilic regions calculate at the same 8 energy levels as W1−W8.
	PSA, PSAR	The Polar Surface Area (PSA) is calculated via the sum of polar GRID atom type contributions, while PSAR is the ratio between the polar surface area (PSA) and the Surface (S).
hydrophobic regions	D1-D8	Volumes of the DRY molecule interaction fields at eight different energy levels, which have been adapted to the energy range of the DRY probe (−0.2 −0.4 −0.6 −0.8 −1.0 −1.2 −1.4 −1.6).
	CD1-CD8	Ratio between the hydrophobic surface and the total molecular surface. It represents the hydrophobic surface per surface unit. Capacity factors are calculated at eight different energy levels, the same levels used to compute the hydrophobic volumes.
	HSA, PHSAR	The Hydrophobic Surface Area (HSA) is calculated via the sum of hydrophobic GRID atom types contributions. The corresponding PHSAR represents the ratio between the polar surface area (PSA) and the hydrophobic surface area.
mixed	FLEX	The Flex descriptor represents the maximum flexibility of a molecule. It is the result of the average of the differences between the maximum and minimum distance of every atom with the others searched on 50 random conformers.
	pharmacophore	These parameters represent 3D pharmacophoric descriptors based on the TOPP (Triplets Of Pharmacophoric Points) descriptors.[18] At first the atoms (points) of a structure are classified as Dry, H-bond donor, and H-bond acceptor, and then all possible triplet of distances between these atoms are generated. The VolSurf+ 3D pharmacophoric descriptors are the maximum conformational area of the triangles derived from every following class of pharmacophoric triplets: DRDRDR (Dry-Dry-Dry), DRDRAC (Dry-Dry-Acceptor), DRDRDO (Dry-Dry-Donor), DRACAC (Dry-Acceptor-Acceptor), DRACDO (Dry-Acceptor-Donor), DRDODO (Dry-Donor-Donor), ACACAC (Acceptor-Acceptor-Acceptor), ACACDO (Acceptor-Acceptor-Donor), ACDODO (Acceptor-Donor-Donor), DODODO (Donor-Donor-Donor).

in the training set compounds and cannot be applied, as it is, for predicting the activity of new compounds with R-groups beyond those used in the analysis.

The classical Free-Wilson model is expressed by the following equation

$$\mathrm{BioActivity} = \sum_{\mathrm{ij}} \alpha_{ij} {*} R_{ij} + \mu \qquad \text{(iii)}$$

where the constant term μ (activity value of the unsubstituted compound) is the overall average of biological activities, and α_{ij} is the R-group contribution of substituent R_i in position j. If substituent R_i is in position j, then $R_{ij} = 1$, otherwise $R_{ij} = 0$. This gives rise to a set of equations that can be potentially solved by MLR, where α_{ij} are the regression coefficients, R_{ij} are the independent variables, and μ is the intercept. Unfortunately MLR cannot be applied directly to the resulting structural matrix due to a linear dependence on its columns.[28] One way to get around these dependencies is to use the Fujita-Ban approach which relates all the structures to a reference structure.[25]

Kubinyi et al. have shown that the original Free-Wilson and the Fujita-Ban modification are linearly related, with the latter approach being a linear transformation of the classical Free-Wilson model.[28] Additionally, the Fujita-Ban model leads to a number of important advantages. First, no complex transformation of the structural matrix is required, and only the removal of one column for each site of substitution is necessary to move from the structural matrix to the Fujita-Ban matrix. Second, the matrix is not changed by the addition or elimination of a compound. Third, in the Fujita-Ban model the constant term μ in the linear equation is derived theoretically by applying the least-squares method and therefore not markedly influenced by the addition or elimination of a compound. In consideration of these advantages the Fujita-Ban methodology was implemented for the analysis reported here.

D. Data Preparation. Compounds with correlated R-groups and outlier compounds whose R-groups did not occur in other compounds were removed from the data set as the activity contribution for these R-groups could not be estimated. Then the remaining structural matrix was rearranged into independent blocks where R-groups from one block would not cross over with other blocks (Figure 2), and statistical analysis was applied to each block separately to estimate the activity contribution for each R-group. Furthermore, blocks whose R-group activity contributions could not be estimated due to a lack in R-group crossovers were further eliminated. This block separation and compound removal procedure maximized the total number of R-group activity contributions that could be estimated.

E. Statistical Analysis. The relationship between the kinase selectivity data and the chemical structures was analyzed using MLR, a multivariate statistical method able to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to the observed data. An MLR model was first built independently for each kinase.

The models predictivity was evaluated by using both "internal validation", by means of standard Leave-One-Out (LOO) analysis, and "external validation". LOO is a cross-validation procedure that works by building reduced models (models for which one object at a time is removed) and using them to predict the Y-variables of the object held out. A more rigorous way to evaluate the reliability of any QSAR model makes use of an external test set; in this case, the predicted activity for the external compounds is compared with the corresponding experimental value.

In the end, the quality of both the internal and external validation was assessed by computing the squared Pearson correlation coefficient (r^2_{corr}) between predicted and actual activities and the associated standard error of correlation (*STE*)

$$r^2_{corr} = \frac{\left[\sum_{i\in test} (y_i^{pred} - \bar{y}_i^{pred})(y_i^{act} - \bar{y}_i^{act})\right]^2}{\sum_{i\in test} (y_i^{pred} - \bar{y}_i^{pred})^2 \sum_{i\in test} (y_i^{act} - \bar{y}_i^{act})^2} \qquad STE = \sqrt{\frac{1}{n-2}\left[\sum_{i\in test} [y_i^{pred} - \bar{y}_i^{pred}]^2 - \frac{\left[\sum_{i\in test} (y_i^{pred} - \bar{y}_i^{pred})(y_i^{act} - \bar{y}_i^{act})\right]^2}{\sum_{i\in test} (y_i^{act} - \bar{y}_i^{act})^2}\right]} \quad \text{(iv)}$$

where y_i^{pred} is the predicted activity for the i^{th} test set compound, y_i^{act} is its measured activity, y_i^{pred} and $^{-}y_i^{act}$ are the average of the predicted and measured activity values respectively, and n is the sample size.

F. Prediction Set. After model building and library enumeration, the activity of each enumerated compound against the kinase targets in the KSS panel was computed and used to derive the overall kinase selectivity profile. A flowchart of the implemented procedure is shown in Figure 5. This resulted in 861 and 1764 compounds being predicted for diaminopyrimidine and pyrrolopyrazole, respectively. Of these 2625 chemical entities, 259 (150 diaminopyrimidines and 109 pyrrolopyrazoles) were present in the in-house liquid store, and the prediction set selection was therefore restricted to this subset of chemically available compounds.

To further shrink the number down, two different selection criteria were applied. First, a set was chosen to be *As Selective As Possible* (ASAP) in an attempt to pick compounds showing the most selective virtual profile. For this purpose, the following rules were applied: a) A compound was considered if showing at least one activity data point (F−W predicted value for a compound against a given kinase target) with $\text{pIC}_{50} \geq 7$ ($\text{IC}_{50} \leq 100$ nM). b) All the compounds passed the previous step were assigned a selectivity score. This was done by counting, for a given compound among the 45 kinases in the panel, how many activity data points were present with $\text{pIC}_{50} \geq 5.3$ ($\text{IC}_{50} \leq 5$ μM). c) At last, compounds were selected trying to

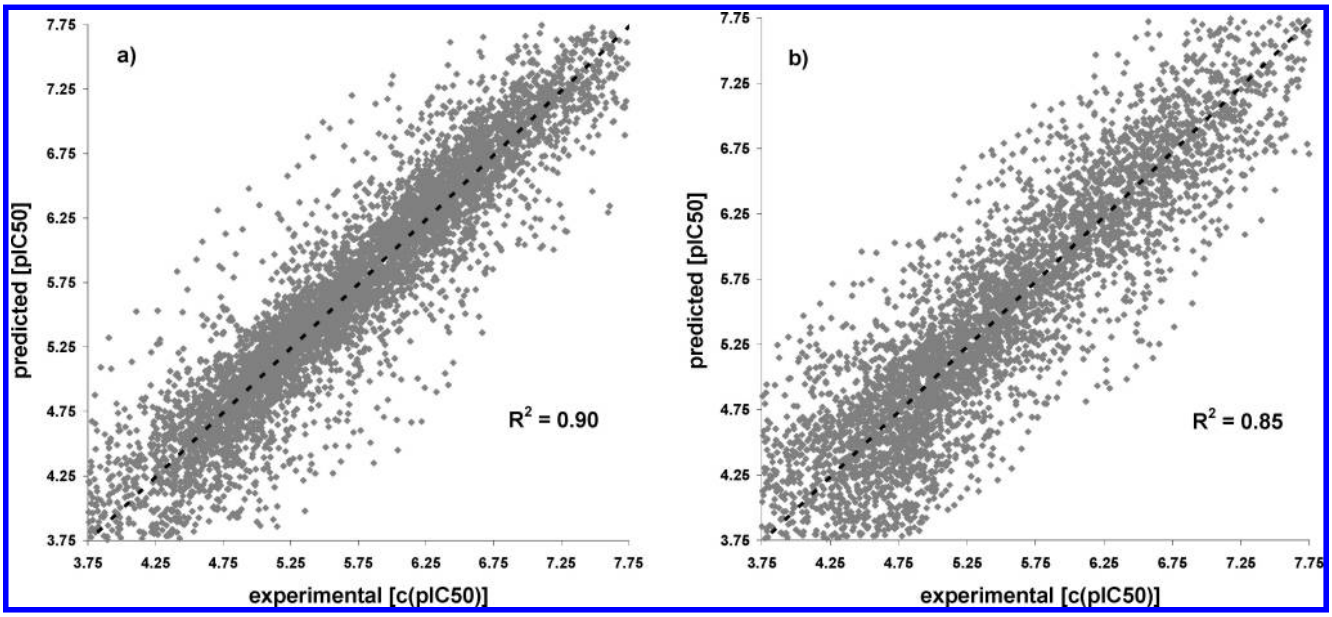


Figure 6. Predicted versus experimental pIC_{50} values for Leave-One-Out (LOO) cross-validation analysis. In general, prediction of pIC_{50} is in good agreement with experimental pIC_{50} (derived from percent inhibition), with a global correlation coefficient $r^2_{\text{corr,CV}} = 0.90$ for the diaminopyrimidine series (a) and $r^2_{\text{corr,CV}} = 0.85$ for the pyrrolopyrazole series (b).

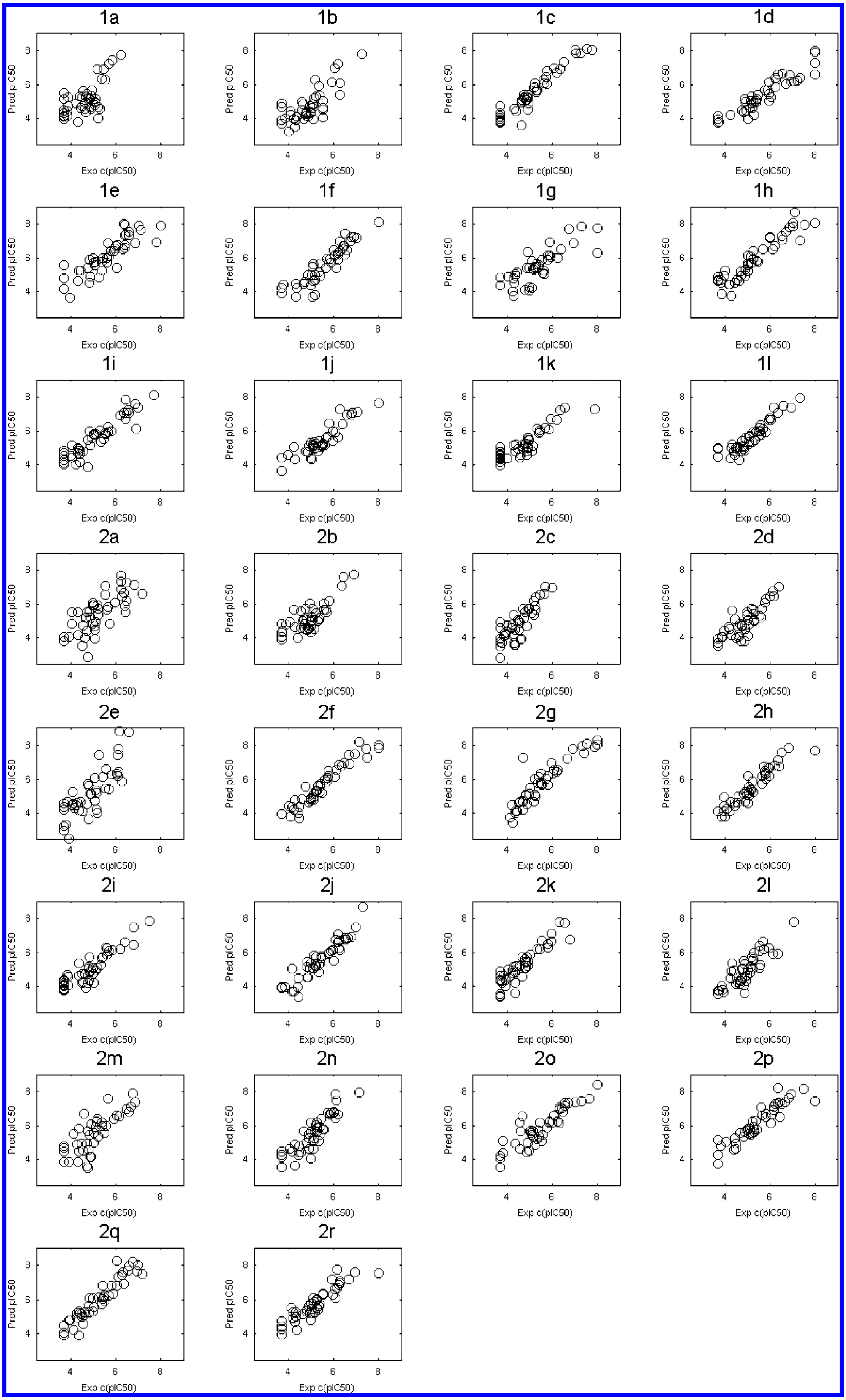


Figure 7. External validation performed on 30 library-enumerated compounds. Each plot shows the correlation between predicted versus experimental pIC_{50} for a given compound across the 45 protein kinases in the KSS panel.

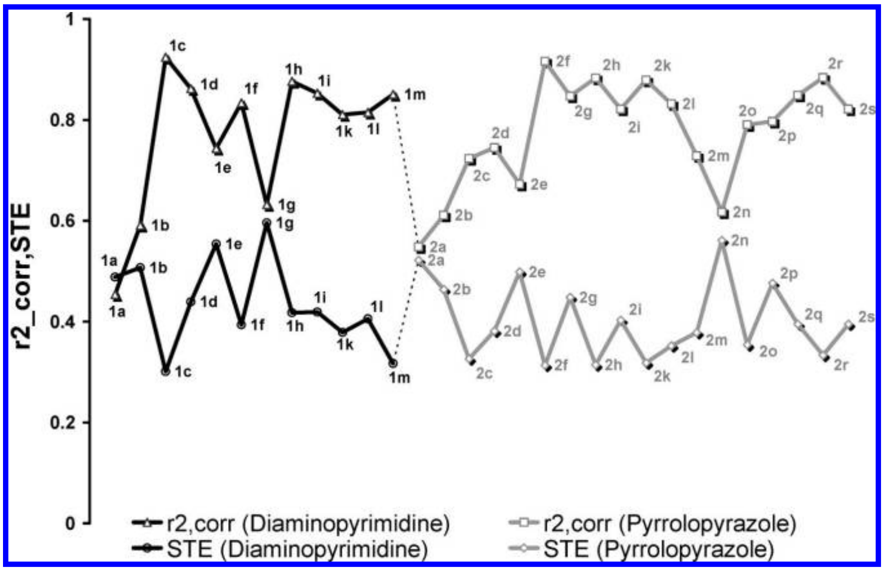


Figure 8. Profile plot showing the squared Pearson coefficient distribution (with the associated standard error of correlation) for the 30 external compounds used in the validation analysis. All compounds were quantitatively well predicted, with an average squared correlation coefficient $r^2_{corr,ext\text{-}pred} = 0.77$ (STE = 0.44) for diaminopyrimidine and $r^2_{corr,ext\text{-}pred} = 0.78$ (STE = 0.41) for pyrrolopyrazole.

maximize both the selectivity score and the diversity in the R-groups. The second set of compounds was selected to be *As Promiscuous As Possible* (APAP). This time we were interested in cherry picking compounds hitting most of the kinase targets in the panel. The ranking scheme was based on a promiscuity score, determined by counting how many activity data points for each compound were present with $pIC_{50} \geq 7$. As before, the final selection was based on trying to maximize both the promiscuity score and the diversity in the R-group structures.

At the end, a total number of 30 compounds, 12 from the diaminopyrimidine core (6 ASAP + 6 APAP) and 18 from the pyrrolopyrazole core (4 ASAP + 14 APAP), were selected as an external test set to be experimentally assessed within the KSS panel (Table 1).

G. ECFP 2D-Fingerprints. Extended Connectivity Fingerprints[39] (ECFPs) are based on a variant of the Morgan algorithm, originally used in isomorphism issues. The generation of these fingerprints begins with the assignment of an initial atom code to each heavy atom present in the molecule based on the following features: the number of connections, the element type, the charge, and the atomic mass. This initial fingerprint is called “ECFP_0” as the maximum diameter explored is only around each atom. An iterative process is used to generate larger structural neighborhoods until the desired size is reached; in our case, the standard ECFP_4. The underlying advantages of this circular abstraction are that the fingerprint calculation is very fast, and the features (size up to 4 billion) are not predefined in a limited fragment dictionary.

H. VolSurf 3D-Descriptors. VolSurf[14,40,41] descriptors are derived from Molecular Interaction Fields (MIF) computed with the program GRID using different probes (in our case water, oxygen, and the hydrophobic probe). The large body of information contained in these fields, on the order of 100,000 grid points, is then analyzed and encoded into physicochemically relevant descriptors. This results in a small number of variables describing the overall distribution of hydrophobic and hydrophilic regions around the molecule (Table 2). Indeed, the molecular descriptors obtained refer to molecular size and shape, both hydrophilic and hydrophobic region size and shape, and to the balance between them. Hydrogen bonding, amphiphilic moments, and critical packing parameters are other useful descriptors. VolSurf descriptors have been shown useful in describing pharmacokinetic and physicochemical properties.[14]

I. R-Group Coefficient Analysis. The objective of this analysis was to gain knowledge from the R-group contributions as determined by the Free-Wilson methodology. Only R-groups for which a coefficient could be determined across the 45 kinases in the panel were taken into account. For the diaminopyrimidine series this resulted in 36 R1- and 26 R2-group structures, giving rise to two different matrices containing 36 × 45 R1- and 26 × 45 R2-group contributions. In the pyrrolopyrazole series, a total of 60 R1 and 35 R2-group structures were available for analysis, leading to two coefficient matrices of 60 × 45 R1- and 35 × 45 R2-group contributions. Position R3 of diaminopyrimidine was not included in this analysis because the coefficients of the R-group structures tested at this position could not be determined across all kinases in the panel. Two different analyses were carried out in order to study the R-group information contained in each of the coefficient matrices: 1) clustering analysis, which consists of computing the similarity metric (using Euclidean distance) between all of the R-group coefficient profiles and then using the Ward[42] hierarchical clustering algorithm, and 2) R-group selectivity analysis, to detect whether small changes in structure could give rise to large variations in activity. This was achieved by computing all pairwise structural similarities between R-groups (using ECFP4 structural descriptors and Tanimoto as similarity measure) and then keeping only the R-group pairs with Tanimoto similarity greater than 0.8. Afterward, each surviving R-group pair was assigned a profile resulting from the difference in the original coefficients profiles for the R-groups being compared.

L. Core Docking. The crystal structures of 21 out of 45 protein kinases present in the KSS were available in the RCSB[43] Protein Data Bank. For these, a core-docking analysis was carried out in order to determine the predicted

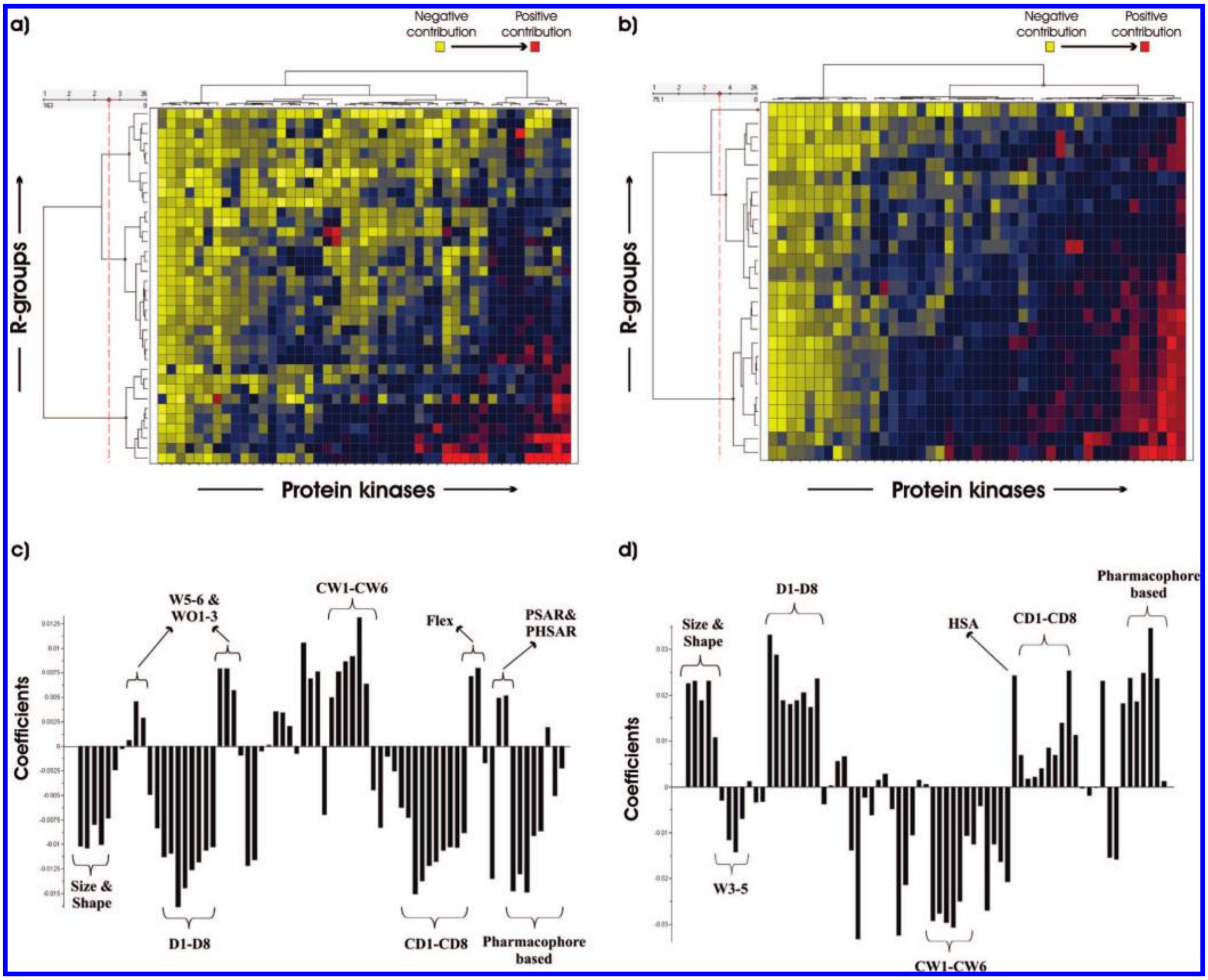


Figure 9. R-group clustering analysis of diaminopyrimidine series for both R1 (a) and R2 (b) positions. a-b) Two-way clustering of kinases and R-groups. Each spot in the heat map represents the R-group activity contribution against a given kinase in the panel as derived by Free-Wilson analysis. Kinases are clustered on the X-axis, while R-groups are clustered along the Y-axis. c-d) PLS coefficients plots using VolSurf descriptors at positions R1 (c) and R2 (d). Positive coefficients directly correlate with an increase in promiscuity, while descriptors whose coefficients are negative are supposed to be relevant for selectivity across different kinases. The most significant descriptors (highly negative and positive) responsible for R-groups/kinase selectivity/promiscuity are highlighted and discussed in the text (see the “R-Group Clustering Analysis” section).

binding mode of each virtual library’s compound with respect to the protein being studied. Here for a virtual compound we refer to a compound whose combination of R-groups was neither originally present in the training set nor tested in the KSS panel.

The core-docking procedure[44] is a protocol specifically designed for screening multiple combinatorial libraries against a family of proteins, in our case the protein kinase family. It relies on a common alignment of all the protein kinase X-ray structures internally available and consists of the following main stages. To begin, each core shown in Figure 1 was converted to SMARTS[45] notation and then used to query a database of all ligands extracted from aligned kinase cocrystal structures. This identified over 50 matches for both cores. Core within a 2 Å threshold were then grouped together yielding 1 unique core position for the pyrrolopyrazole core and 2 unique core positions for the diaminopyrimidine core. These experimentally determined core position references were then used to overlay all the virtual compounds, and the AgDock[46,47] docking software was applied to optimize their overall position within the binding site of each protein kinase under investigation, with the AgDock developed empirical scoring function (HTscore),[48] utilized to select one docked pose with the best score for each input ligand.

RESULTS AND DISCUSSION

Model Building and Validation. In this section we discuss the QSAR modeling results obtained by applying Free-Wilson to the structural matrix of descriptors for a set of diaminopyrimidine and pyrrolopyrazole inhibitors. After compounds in both series were fragmented, based on the core schemes given in Figure 1, the corresponding structural matrix for each protein kinase was built, including every compound for which the inhibition was experimentally determined in the KSS panel.

Statistical analysis was performed using MLR with the squared Pearson correlation coefficients for diaminopyrimidine and pyrrolopyrazole series across the 45 protein kinases

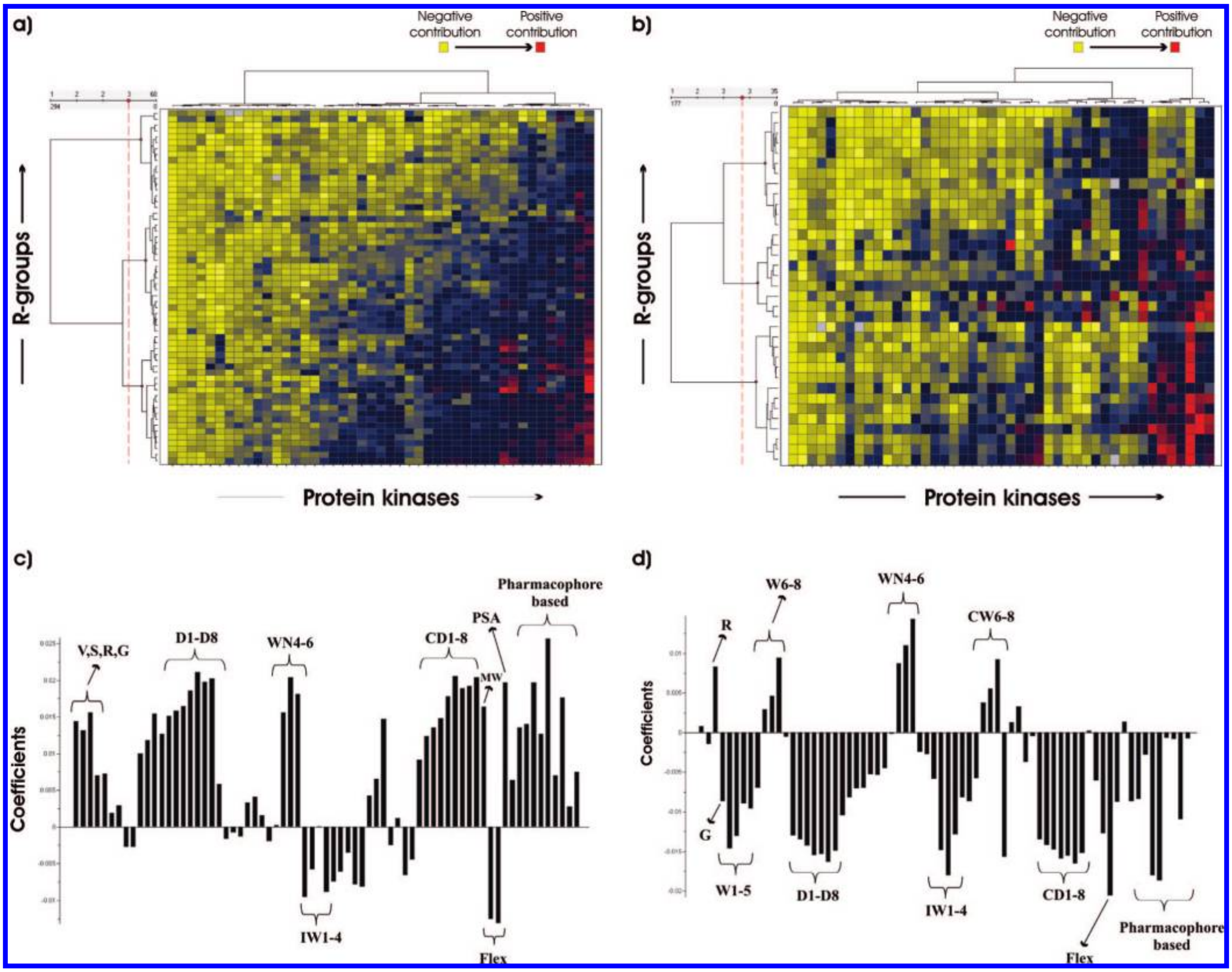


Figure 10. R-group clustering analysis of pyrrolopyrazole series for both R1 (a) and R2 (b) positions. a-b) Two-way clustering of kinases and R-groups. Each spot in the heat map represents the R-group activity contribution against a given kinase in the panel as derived by Free-Wilson analysis. Kinases are clustered on the X-axis, while R-groups are clustered along the Y-axis. c-d) PLS coefficients plots using VolSurf descriptors at position R1 (c) and R2 (d). Positive coefficients directly correlate with an increase in promiscuity, while descriptors whose coefficients are negative are supposed to be relevant for selectivity across different kinases. The most significant descriptors (highly negative and positive) responsible for R-groups/kinase selectivity/promiscuity are highlighted and discussed in the text (see the "R-Group Clustering Analysis" section).

being respectively in the range of r^2_{fitting}=0.82÷0.95 (average r^2_{fitting}=0.87) and r^2_{fitting}=0.73÷0.93 (average r^2_{fitting}=0.85). To further validate the models, a LOO validation procedure was carried out, and the results are shown in Figure 6. In general, our predicted pIC_{50} is in agreement with the pIC_{50} derived from percent inhibition values. In the diaminopyrimidine series, taking all 45 models together, 6712 LOO estimations were carried out giving a global correlation coefficient $r^2_{\text{corr,CV}} = 0.90$ and a standard error of the predicted pIC_{50} value in the regression STE = 0.35, with about 88% of all predictions made within ± 0.5 error (Figure 6a). Similar results were obtained for the pyrrolopyrazole series, where LOO estimations of 5413 objects gave an overall correlation coefficient $r^2_{\text{corr,CV}} = 0.84$ and a standard error of the predicted pIC_{50} value in the regression STE = 0.47, with about 75% of the overall predictions falling within ± 0.5 error (Figure 6b).

Since Free-Wilson models use the presence or absence of distinct R-group fragments as the basic variables in regression, the derived model coefficients can be treated as a quantitative estimate of the activity contribution of each R-group. Assuming the additive assumption holds, then these R-group contributions can be used to make reliable predictions for all the enumerated compounds in a virtual library, where all R1 fragments are crossed with all R2 fragments.

To test this hypothesis, 30 enumerated compounds within the virtual libraries (12 diaminopyrimidine and 18 pyrrolopyrazole compounds), available in our liquid store, and defined by the monomers of the training set were selected and submitted for testing at both 1 and 10 μM in KSS (see Computational Details). The resulting values were transformed into calculated pIC_{50} applying eq ii and then plotted versus predicted pIC_{50}, derived from Free-Wilson estimations of R-groups activity contributions. The predictions for the test set are graphically represented in Figure 7, where each box shows the correlation between predicted and calculated pIC_{50} across all 45 protein kinases in the panel. All compound profiles were quantitatively well predicted, with an overall squared correlation coefficient $r^2_{\text{corr,ext-pred}} = 0.77$ (STE = 0.44) for diaminopyrimidine and $r^2_{\text{corr,ext-pred}} = 0.78$

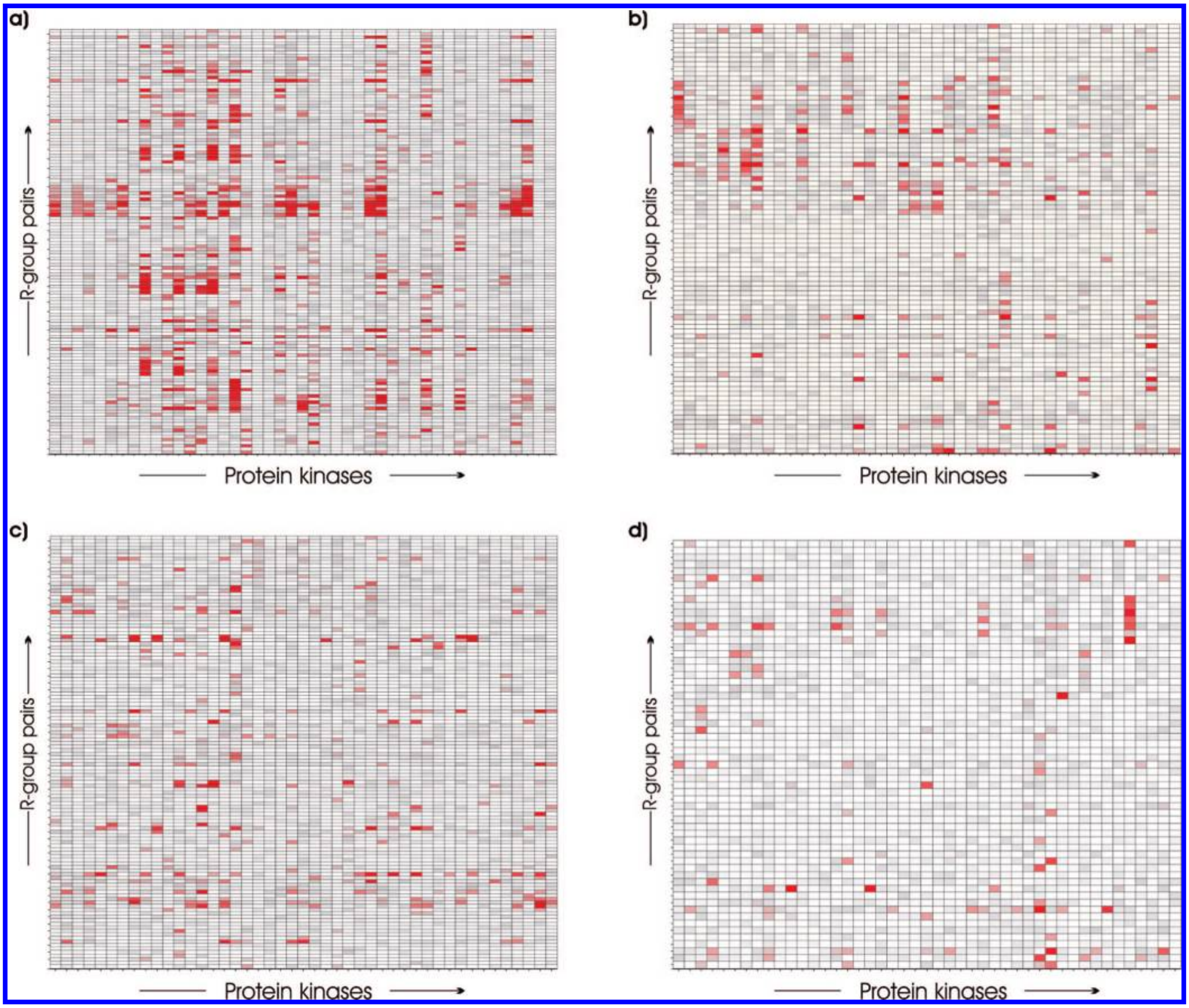


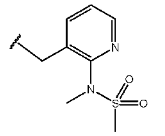
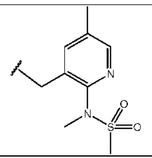
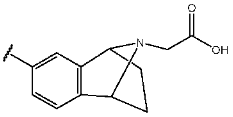
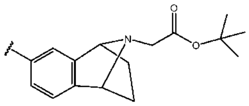
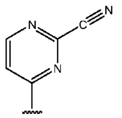
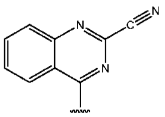
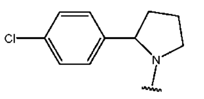
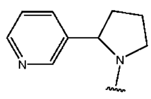
Figure 11. R-group selectivity maps. Only pairs of R-groups with Tanimoto similarity greater than 0.8 are reported on the Y-axis, while all 45 proteins kinases are listed on the X-axis. Each spot in the heat maps represents the combination of an R-group pair with a given protein kinase. Red spots indicate interesting combinations where small changes in the structure could give rise to large variation in activity. Combinations are assigned a color ranging from white ($\Delta pIC_{50} = 0$) to red ($\Delta pIC_{50} \geq 2$).

(STE = 0.41) for pyrrolopyrazole (Figure 8), showing how the Free-Wilson methodology can be efficiently applied in QSAR studies where the additive assumption seems to be satisfied, as in our data set.

Virtual Library Space Analysis. Due to experimental and synthetic limitations, typically only a small number of compounds can be synthesized and screened against a kinase selectivity panel. As a result, many compounds with desired selectivity profiles could potentially be missed. By using high-quality QSAR models, the selectivity profiles of compounds from the entire virtual library can be reliably estimated, thus, greatly expanding the chemical space coverage and increasing the chance of finding compounds with attractive profiles. To demonstrate this, we enumerated the full virtual library for the two series. We obtained 861 compounds for the diaminopyrimidine series and 1764 compounds for the pyrrolopyrazole series, using R-groups of the 321 existing compounds (195 diaminopyrimidine, 126 pyrrolopyrazole). We then calculated their selectivity profile using the QSAR models derived from Free-Wilson analysis. Among the existing compounds, 18 of them (17 diaminopyrimidines, 1 pyrrolopyrazole) met our selectivity criteria ($pIC_{50} > 5.3$ against no more than 5 kinases on the panel). In the full virtual library, however, 65 additional compounds (57 diaminopyrimidines, 8 pyrrolopyrazoles) were predicted to be selective.

The expansion of the inhibitor's selectivity space leads to an increase in the number of kinases selectively targeted, suggesting that such a procedure would also be suitable as a tool for exploring potential "Target Hopping". Indeed, when applied to our data set, existing selective compounds from the diaminopyrimidine and the pyrrolopyrazole series targeted 14 and 5 protein kinases, respectively. However, after complete enumeration of the virtual libraries, 28 and 19 protein kinases were predicted to be selectively inhibited by compounds in the two series respectively. This shows how series originally developed for a specific kinase could be turned into selective inhibitors for other kinases by exploiting different R-group combinations.

Table 3. R-Group/Kinase Contributions from Free-Wilson Selectivity Maps

Core	SITE	R-GROUP[A]	R-GROUP[B]	PKS	F-W $^{\Delta pIC50}$ $(R^2 - R^1)$
Diaminopyrimidine	R1			**GSK3β** (1O9U)	+1.8
	R2			**PAK4** (2CDZ)	+2.5
Pyrrolopyrazole	R1			**NEK2** (2CL1)	+1.6
	R2			**PDK1** (1Z5M)	+2.1

R-Group Clustering Analysis. An important feature of any QSAR procedure is the identification of which structural properties have the largest affect on the biological activity of a given chemical series.

To accomplish this task a combination of R-group coefficients clustering followed by molecular descriptors based linear regression was applied to study possible global trends and patterns between R-group structural similarities and the propensity to generate certain kinase selectivity profiles.

In the diaminopyrimidine series, Ward hierarchical clustering was used to analyze both R1 and R2 position coefficient matrices, containing 36 × 45 and 26 × 45 R-group contributions, respectively. This hierarchical clustering procedure sorts the data in two dimensions, bringing kinases and R-groups with similar profiles together. Results of the cluster analysis are reported as heat maps in Figure 9, where each R-group contribution is assigned a color ranging from yellow (negative influence on the activity) to red (positive influence on the activity). These heat maps also describe an increasing tendency, from top to bottom, for an R-group to become less selective.

To identify the most relevant physicochemical properties driving this trend in R-groups/kinase selectivity, VolSurf three-dimensional descriptors were generated for the R1 and R2 position of the diaminopyrimidine. Partial Least-Squares (PLS) statistical analysis was applied using the hierarchical clustering ordering function (which describes the vertical order in the row dendrogram) as a dependent variable. Coefficient plots for R1 and R2 showing the contribution of VolSurf descriptors to explain the hierarchical clustering results are reported in Figure 9. Variables with the highest positive and negative values are considered to be the most significant to explain the selectivity trend in our data set, with the positive variables directly correlating to an increase in promiscuity and the negative variables describing VolSurf descriptors important for selectivity across different kinases.

In the case of R1, descriptors related to the size and shape of the molecule such as molecular surface (S), volume (V), and globularity (G) (defined as ratio between the surface (S) and the surface of a sphere with the same volume (V)), together with wide hydrophobic regions (descriptors D1-D8; CD1-CD2, HSA) and specific pharmacophores (Dry-Dry-Dry, Dry-Dry-HBA, Dry-Dry-HBD) were inversely correlated with kinase promiscuity, indicating that an increase for these R-group properties leads to an enhancement in kinase selectivity across the panel. Conversely, R-group promiscuity seems more related to wide hydrophilic regions (W5−6, CW1−6, IW1−3), flexibility (FLEX), and polar surface area descriptors (PSAR, PHSAR) (Figure 9).

A different and opposite trend in physicochemical properties drives R-group selectivity at the R2 position of the diaminopyrimidine core. Here promiscuity is directly linked with surface, volume, globularity, and, what is more interesting, with hydrophobic based VolSurf descriptors (D1−8, CD1−8, HSA) and pharmacophores based features (Dry-HBA-HBD being the most significant), indicating the presence of a bulky hydrophobic subpocket with a potential hydrogen bond pattern common across different kinases (Figure 9).

The same analysis has been performed in the case of the pyrrolopyrazole series, with Ward hierarchical clustering used for both R1 and R2 positions, consisting of two coefficient matrices containing 60 × 45 and 35 × 45 R-group contributions, respectively. The resulting output, reported as dendrograms in Figure 10, is again sorted based on the hierarchical cluster order function, with an increasing tendency from top to bottom for an R-group to become more promiscuous. Coefficient plots, which show the contribution of VolSurf descriptors to explain the results from hierarchical clustering, are also reported in Figure 10. From the R1 coefficient plot, descriptors related to the size and shape of the molecule such as molecular surface (S), volume (V), globularity (G), and rugosity (degree of the molecular wrinkled surface), together with wide hydrophobic regions (descriptors D1-D8; CD1-CD2, HSA) and pharmacophores based features arrangements (Dry-Dry-HBD being the most relevant) show high positive values, indicating that an increase in these R-group properties leads to an enhancement

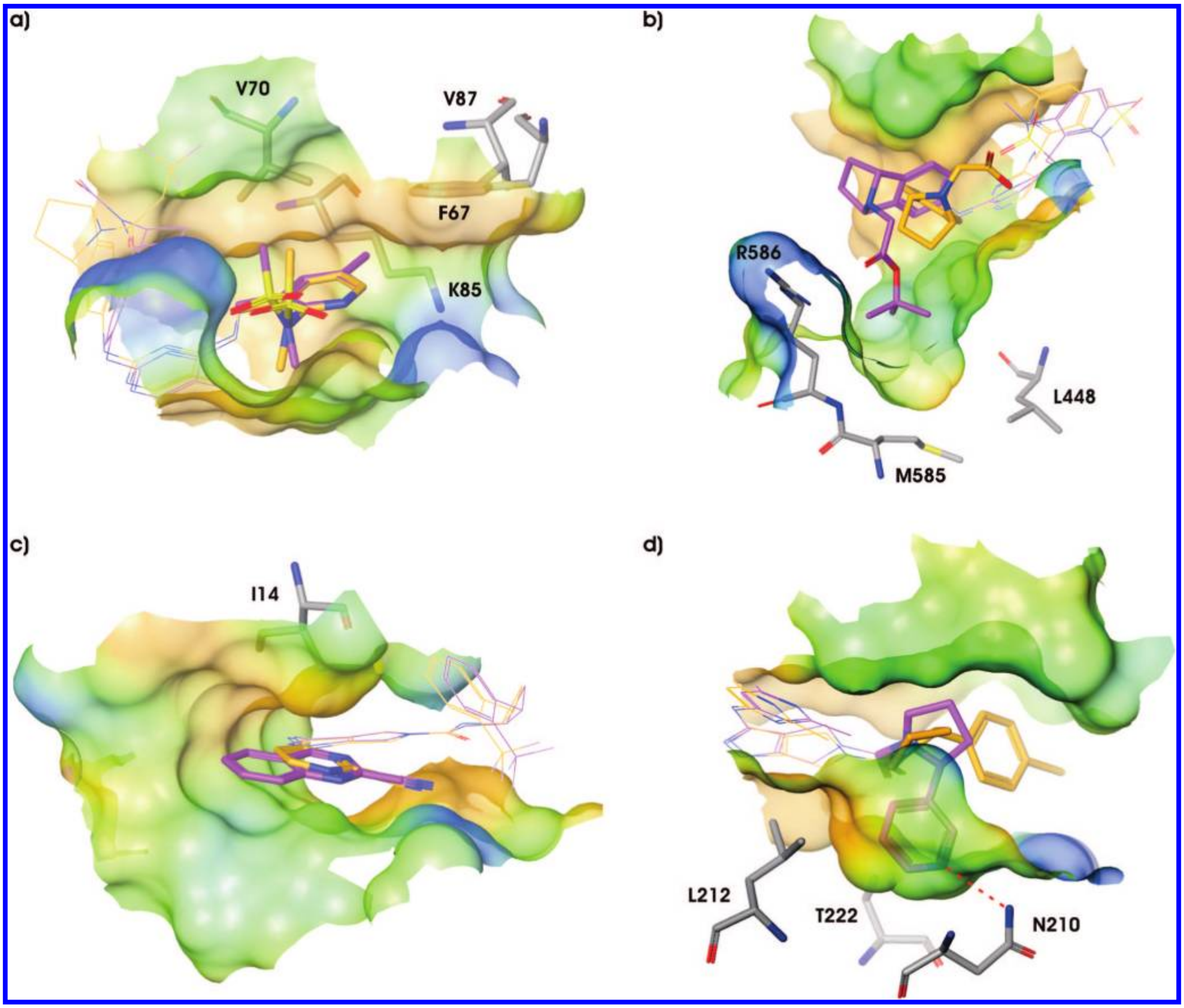


Figure 12. Structural models for binding site interactions of diaminopyrimidine and pyrrolopyrazole series. a) R-group^A (orange) and R-group^B (violet) at site R1 of diaminopyrimidine docked into the crystal structure of GSK3β (1O9U). The extra methyl in R-group^B is responsible for its increased activity contribution. b) Position R2 of diaminopyrimidine in protein kinase PAK4 (2CDZ). R-group^B (violet) undergoes a 45° rotation in order to orient the tert-butyloxy tail toward the buried lipophilic pocket. c) Docking model for protein kinase NEK2 (2CL1) explaining the raise in R-group activity contribution when moving from R-group^A (orange) to R-group^B (violet) at the R1 position in pyrrolopyrazole, with the structural rationalization being the extended hydrophobic interaction with residue I14 in the binding site. d) Site R2 in pyrrolopyrazole. Variation in R-group composition determines a different binding mode for R-group^A (orange) compared to R-group^B (violet), with the latter filling a lipophilic pocket in the active site of PDK1 (1Z5M) and making a hydrogen bond interaction with residue N210.

of kinase promiscuity across the panel. However, the high coefficient values for PSA (positive) and IW1−4 (negative, showing the unbalance between the center of mass of a molecule and the barycenter of its hydrophilic regions) indicate that hydrophilic regions can also play an important role and that an appropriate hydrophobic-hydrophilic balance may be a prerequisite for an R-group to bind at this position of the pyrrolopyrazole series. Another important descriptor able to differentiate R-groups within the R1 dendrogram was molecular flexibility, the most selective compounds being the most flexible.

As seen in the diaminopyrimidine series, position R2 of pyrrolopyrazole is driven by different global physicochemical properties than R1. The most relevant size and shape descriptor was rugosity, while volume, surface, and molecular weight were not important clustering properties. Hydrophobic descriptors correlated negatively to the overall PLS model (D1−8, CD1−8); indeed, the most selective R-groups were also the most hydrophobic, with part of the structure being flat and aromatic, in contrast with R-groups clustered at the bottom of the dendrogram (less selective) where an increasing aliphatic character was seen.

Other significant contributions to the model came from a subset of hydrophilic descriptors (W1−8, WN4−6, CW6−8, IW1−4), indicative of the fact that an appropriate hydrophobic-hydrophilic balance is also an important prerequisite for R-group binding, confirmed by the high coefficients obtained for the pharmacophoric based descriptors Dry-Dry-HBD and Dry-HBA-HBD. All together, the coefficients profiled at position R2 of pyrrolopyrazole pointed out the possible presence of a flat and hydrophobic subpocket with a critical geometrical arrangement of hydrogen bond features that could potentially be used for mining R-groups/kinase selectivity across the panel.

R-Group Selectivity Analysis. With the aim to further validate the results from the QSAR models, based on the availability of numerous public protein kinase crystal structures, we combined the information from the Free-Wilson approach with the 3D structural knowledge of the target, to provide more insight for kinase selectivity.

First, the R-group coefficient maps were transformed into selectivity maps, allowing the detection of small changes in structure that could give rise to large variation in activity (see Computational Details). Figure 11 shows the results of this transformation, reported as heat maps where each R-group pair/kinase combination is assigned a color ranging from white (difference in R-group contribution, $\Delta pIC_{50} = 0$) to red ($\Delta pIC_{50} \geq 2$).

Finally, a structure-based study was carried out for each R-group and protein kinase combination using the internal core-docking routine. An exhaustive structure-based interpretation of the R-group contributions shown in the selectivity heat maps of Figure 11 is beyond the scope of this study. Our objective was spot checking the ligand-based results obtained through the Free-Wilson analysis to see if they made sense in the context of the known kinase structure; therefore, only one example for each site of substitution for both diaminopyrimidine and pyrrolopyrazole is shown here (Figure 11a-d and Table 3).

Starting with the R1 position of diaminopyrimidine, both poses for R-group^A and R-group^B, as reported in Table 3, were analyzed after docking into the protein kinase GSK3β subfamily (PDB entry: 1O9U). A variation in pIC_{50} of 1.8 logarithmic units was found using Free-Wilson calculations for estimating the activity contributions of these R-groups. The only structural difference between the two is a methyl in position 5 of the pyridine ring. Although the docking study showed the same binding mode, the methyl moiety in R-group^B is now buried into the protein kinase active site and pointing toward a small lipophilic pocket (F67, V70, K85, V87), explaining the increase in activity predicted by the Free-Wilson model (Figure 12a).

A different combination of R-groups/protein kinase was examined using the R2 position of diaminopyrimidine (Figure 11 and Table 3). Figure 12b shows the resulting poses for R-group^A and R-group^B (Table 3) when docked into the PAK4 protein kinase binding site (PDB entry: 2CDZ). Changing from the carboxy- to the tert-butyloxy-moiety forces a different binding orientation of the R-groups within the active site. The structure-based rationalization for pIC_{50} difference ($\Delta pIC_{50} = 2.5$) is R-group^B which undergoes a 45° rotation, around the C−N single bond linking the R-group to the diaminopyrimidine core, allowing the tert-butyloxy tail to orient in the direction of a buried lipophilic pocket made by cavity flanking residues L448, M585, and R586 (Figure 12b).

Similar conclusions can be derived when analyzing the core-docking results for the pyrrolopyrazole series. Protein kinase NEK2 (PDB entry: 2CL1) was chosen to elucidate the rise in activity ($\Delta pIC_{50} = 1.6$) when moving from pyrimidine-2-carbonitrile to quinazoline-2-carbonitrile at the R1 position of the pyrrolopyrazole core. Figure 12c highlights the structural explanation for that, where the presence of the second 6-membered aromatic ring of quinazoline is not influencing the R-group binding mode but is extending the staked hydrophobic interaction toward residue I14.

The last example examines the R2 position of pyrrolopyrazole where large differences in pIC_{50} ($\Delta pIC_{50} = 2.1$) were achieved by substituting two highly similar R-groups (chlorobenzene vs pyridine), when tested against protein kinase PDK1 (PDB entry: 1Z5M). The rationale for this can be found by looking at the core-docking results, as shown in Figure 12d. Variation in R-group composition determines a different binding mode for the two R-groups, with the pyridine portion of R-group^B filling a small pocket in the active site made up of a combination of lipophilic (L212, T222) and hydrophilic (N210) residues. Indeed, a pyridine moiety at this position is capable of establishing favorable hydrogen bond interactions with the nitrogen of N210, while maximizing few hydrophobic contacts with the surrounding L212 and T222 residues.

SUMMARY AND PERSPECTIVES

Although drug selectivity is not always necessary for activity (i.e., Gleevec, whose primary intended target was ABL but also inhibits the c-KIT and PDGF-receptor tyrosine kinases with similar potency),[49] off-target effects and associated toxicity is a serious and common problem in the inhibition of kinase function.

In the investigation presented here, Free-Wilson analysis was initially applied to the modeling of 700 kinase inhibitors, belonging to two different chemical series, screened against a panel of 45 structurally diverse protein kinases. Statistically significant QSAR models were obtained as well as overall agreement between predicted vs experimental pIC_{50} was achieved for both diaminopyrimidine and pyrrolopyrazole series.

The predictive power of our QSAR models was verified computationally in LOO cross validation study and experimentally using external test compounds. Modeling the selectivity panel data with the Free-Wilson method enables scientists to quickly assess the selectivity profile for the entire virtual library space and quickly identify compounds with desirable selectivity profiles. In comparison with other QSAR modeling methods, the Free-Wilson method has the advantage of generating activity contribution estimates for individual R-group structures that are readily interpretable to medicinal chemists. The use of R-groups as descriptors in model building gives the models a well defined boundary of the chemical space that can be predicted.

The construction of R-group selectivity profiles based on the estimated R-group contributions against each kinase on the selectivity panel enables us to identify selectivity trends among the R-groups and the physicochemical properties attributed to these trends. R-group selectivity profiles also enable us to systematically identify structural determinants for selectivity where small modification in the R-groups results in a significant difference in selective profiles. The structural knowledge obtained here provides substrates for scientists to formulate novel lead transformation ideas for kinase inhibitors with better quality.

Finally, the availability of X-ray data for many of the kinases structures utilized in the KSS panel allowed us to structurally validate the R-group's Free-Wilson contributions to kinase activity using a ligand docking procedure, confirming the effectiveness of the ligand-based QSAR results.

In our validation study, this approach has proven to be a successful strategy in the quest for selective kinase inhibitors, a critical point in the realization of the "chemogenomics" concept,[50] where targets are no longer viewed as individual and single entities but grouped into sets of related proteins or target families (i.e., kinases, PDEs, MMPs, GPCRs) that are systematically explored. Indeed, the existing procedure can be potentially extended to the study of any set of targets where collections of large libraries of compounds obtained through combinatorial chemistry have been routinely screened against a panel of proteins belonging to the same superfamily.

ACKNOWLEDGMENT

Pfizer, Inc. is acknowledged for continuing support of this research. Jaroslaw Kostrowicki and Paul Rejto are acknowledged for the Java component of the core Free-Wilson calculation. The authors gratefully thank Molecular Discovery for providing the VolSurf program. S. Sciabola expresses personal gratitude to Prof. Gabriele Cruciani and Alexis C. Meng for helpful discussions during the preparation and revision of this paper.

REFERENCES AND NOTES

(1) Manning, G.; Whyte, D. B.; Martinez, R.; Hunter, T.; Sudarsanam, S. The Protein Kinase Complement of the Human Genome. *Science* **2002**, *298*, 1912–1934.

(2) Kostich, M.; English, J.; Madison, V.; Gheyas, F.; Wang, L.; Qiu, P.; Greene, J.; Laz, T. M. Human members of the eukaryotic protein kinase family. *Genome Biol.* **2002**, *3*, 0043.10043.12.

(3) Johnson, L. N.; Lewis, R. J. Structural Basis for Control by Phosphorylation. *Chem. Rev.* **2001**, *101*, 2209–2242.

(4) Nagar, B.; Bornmann, W. G.; Pellicena, P.; Schindler, T.; Veach, D. R.; Miller, W. T.; Clarkson, B.; Kuriyan, J. Crystal Structures of the Kinase Domain of c-Abl in Complex with the Small Molecule Inhibitors PD173955 and Imatinib (STI-571). *Cancer Res.* **2002**, *62*, 4236–4243.

(5) George, S. Sunitinib, a multitargeted tyrosine kinase inhibitor, in the management of gastrointestinal stromal tumor. *Curr. Oncol. Rep.* **2007**, *9*, 323–327.

(6) Yun, C.-H.; Boggon, T. J.; Li, Y.; Woo, M. S.; Greulich, H.; Meyerson, M.; Eck, M. J. Structures of Lung Cancer-Derived EGFR Mutants and Inhibitor Complexes: Mechanism of Activation and Insights into Differential Inhibitor Sensitivity. *Cancer Cell* **2007**, *11*, 217–227.

(7) Stamos, J.; Sliwkowski, M. X.; Eigenbrot, C. Structure of the Epidermal Growth Factor Receptor Kinase Domain Alone and in Complex with a 4-Anilinoquinazoline Inhibitor. *J. Biol. Chem.* **2002**, *277*, 46265–46272.

(8) Fabian, M. A.; Biggs, W. H.; Treiber, D. K.; Atteridge, C. E.; Azimioara, M. D.; Benedetti, M. G.; Carter, T. A.; Ciceri, P.; Edeen, P. T.; Floyd, M.; Ford, J. M.; Galvin, M.; Gerlach, J. L.; Grotzfeld, R. M.; Herrgard, S.; Insko, D. E.; Insko, M. A.; Lai, A. G. L.; Le′lias, J.-M.; Mehta, S. A.; Milanov, Z. V.; Velasco, A. M.; Wodicka, L. M.; Patel, H. K.; Zarrinkar, P. P.; Lockhart, D. J. A small molecule-kinase interaction map for clinical kinase inhibitors. *Nat. Biotechnol.* **2005**, *23*, 329–336.

(9) Card, A.; Caldwell, C.; Min, H.; Lokchander,B.; Xi, H.; Sciabola, S.; Kamath, A. V.; Clugston, S.; Tschantz, W. R.; Wang, L.; Moshinsky, D. J. High-Throughput Biochemical Kinase Selectivity Assays: Panel Development and Screening Applications. *J. Biomol. Screen.* 2008, Submitted for publication.

(10) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Model.* **2002**, *42*, 1273–1280.

(11) Barnard, J. M.; Downs, G. M. Chemical Fragment Generation and Clustering Software. *J. Chem. Inf. Model.* **1997**, *37*, 141–142.

(12) Nilsson, J.; Wikstrom, H.; Smilde, A.; Glase, S.; Pugsley, T.; Cruciani, G.; Pastor, M.; Clementi, S. GRID/GOLPE 3D quantitative structure-activity relationship study on a set of benzamides and naphthamides, with affinity for the dopamine D3 receptor subtype. *J. Med. Chem.* **1997**, *40*, 833–840.

(13) Pastor, M.; Cruciani, G.; McLay, I.; Pickett, S.; Clementi, S. GRid-INdependent Descriptors (GRIND): A novel class of alignment-independent three dimensional molecular descriptors. *J. Med. Chem.* **2000**, *43*, 3233–3243.

(14) Cruciani, G.; Crivori, P.; CArrupt, P. A.; Testa, B. Molecular Fields in Quantitative Structure-Permeation Relationships: The VolSurf Approach. *THEOCHEM* **2000**, *503*, 17–30.

(15) 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*, 5959–5967.

(16) Mason, J. S.; Morize, I.; Menard, P. R.; Cheney, D. L.; Hulme, C.; Labaudiniere, R. F. New 4-point pharmacophore method for molecular similarity and diversity applications: overview of the method and applications, including a novel approach to the design of combinatorial libraries containing privileged substructures. *J. Med. Chem.* **1999**, *42*, 3251–3264.

(17) Baroni, M.; Cruciani, G.; Sciabola, S.; Perruccio, F.; Mason, J. S. A Common Reference Framework for Analyzing/Comparing Proteins and Ligands. Fingerprints for Ligands And Proteins (FLAP): Theory and Application. *J. Chem. Inf. Model.* **2007**, *47*, 279–294.

(18) Sciabola, S.; Morao, I.; deGroot, M. J. Pharmacophoric Fingerprint Method (TOPP) for 3D-QSAR Modeling: Application to CYP2D6 Metabolic Stability. *J. Chem. Inf. Model.* **2007**, *47*, 76–84.

(19) Geladi, P.; Kowalski, B. R. Partial Least-Squares Regression: A Tutorial. *Anal. Chim. Acta* **1986**, *186*, 1–17.

(20) Zupan, J.; Gasteiger, J. *Neural Networks in Chemistry and Drug Design*, 2nd ed.; Wiley-VCH: 1999.

(21) Breiman, L. Random Forests. *Machine Learning* **2001**, *45*, 5–32.

(22) Cortes, C.; Vapnik, V. Support-vector network. *Machine Learning* **1995**, *20*, 273–297.

(23) Boser, B.; Guyon, I.; Vapnik, V. A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*; 1992.

(24) Barnard, J. M.; Downs, G. M.; Willett, P. Descriptor-based similarity measures for screening chemical databases In *Virtual Screening for Bioactive Molecules*; Böhm, H.-J., Schneider, G. Eds.; Wiley-VCH: 2000; Vol. 10, pp 59−80.

(25) Fujita, T.; Ban, T. Structure-Activity Study of Phenethylamines as Substrates of Biosynthetic Enzymes of Sympathetic Transmitters. *J. Med. Chem.* **1971**, *14*, 148–152.

(26) Hernandez-Gallegos, Z.; Lehmann, P. A. A Free-Wilson/Fujita-Ban analysis and prediction of the analgesic potency of some 3-hydroxy- and 3-methoxy-N-alkylmorphinan-6-one opioids. *J. Med. Chem.* **1990**, *33*, 2813–2817.

(27) Kubinyi, H.; Kehrhahn, O. H. Quantitative structure-activity relationships. 1. The modified Free-Wilson approach. *J. Med. Chem.* **1976**, *19*, 578–586.

(28) Kubinyi, H.; Kehrhahn, O. H. Quantitative structure-activity relationships. 3. A comparison of different Free-Wilson models. *J. Med. Chem.* **1976**, *19*, 1040–1049.

(29) Free, S. M.; Wilson, J. W. A Mathematical Contribution to Structure-Activity Studies. *J. Med. Chem.* **1964**, *7*, 395–399.

(30) Craig, P. N. Structure-activity correlations of antimalarial compounds. 1. Free-Wilson analysis of 2-phenylquinoline-4-carbinols. *J. Med. Chem.* **1972**, *15*, 144–149.

(31) Nisato, D.; Wagnon, J.; Callet, G.; Mettefeu, D.; Assens, J. L.; Plouzane, C.; Tonnerre, B.; Pliska, V.; Fauchere, J. L. Renin inhibitors. Free-Wilson and correlation analysis of the inhibitory potency of a series of pepstatin analogs on plasma renin. *J. Med. Chem.* **1987**, *30*, 2287–2291.

(32) Schaad, L. J.; Hess, B. A.; Purcell, W. P.; Cammarata, A.; Franke, R.; Kubinyi, H. Compatibility of the Free-Wilson and Hansch quantitative structure-activity relations. *J. Med. Chem.* **1981**, *24*, 900–901.

(33) Tomic, S.; Nilsson, L.; Wade, R. C. Nuclear Receptor-DNA Binding Specificity: A COMBINE and Free-Wilson QSAR Analysis. *J. Med. Chem.* **2000**, *43*, 1780–1792.

(34) Ekins, S.; Gao, F.; Johnson, D. L.; Kelly, K. G.; Meyer, R. D. Single point interaction screen to predict IC50. EP 1 139 267 A2, 26.03.2001, 2001.

(35) Schnurr, B.; Schächtele, C. Use of FlashPlate for Automated Kinase Assays *Perkin Elmer Application Note FlashPlate® File #6*. www.perkinelmer.com/lifesciences (accessed Oct 18, 2007).

(36) Hastie, C. J.; McLauchlan, H. J.; Cohen, P. Assay of protein kinases using radiolabeled ATP: a protocol. *Nature Protocols* **2006**, *1*, 968–971.

(37) Johnson, M.; Li, C.; Rasnow, B.; Grandsard, P.; Xing, H.; Fields, A. Converting a Protease Assay to a Caliper Format LabChip System. *J. Assoc. Lab. Automat.* **2002**, *7*, 62–68.

(38) Dunne, J.; Reardon, H.; Trinh, V.; Li, E.; Farinas, J. Comparison of On-Chip and Off-Chip Microfluidic Kinase Assay Formats. *Assay Drug Dev. Technol.* **2004**, *2*, 121–129.

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

(40) Crivori, P.; Cruciani, G.; Carrupt, P. A.; Testa, B. Predicting blood-brain barrier permeation from three-dimensional molecular structure. *J. Med. Chem.* **2000**, *43*, 2204–2216.

(41) Cruciani, G.; Pastor, M.; Guba, W. VolSurf: a new tool for the pharmacokinetic optimization of lead compounds. *Eur. J. Pharm. Sci.* **2000**, *11*, S29–39.

(42) Ward, J. H. Hierarchical Grouping to Optimize an Objective Function. *J. Am. Stat. Assoc.* **1963**, *58*, 236.

(43) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.

(44) Wittkopp, S.; Penzotti, J. E.; Stanton, R. V.; Wildman, S. A. Knowledge-based docking for kinases with minimal bias. In *234th ACS National Meeting* Boston, MA, United States, 2007.

(45) Daylight Chemical Information System Inc. 120 Vantis - Aliso Viejo, CA 92656. http://www.daylight.com (accessed July 27, 2007).

(46) Gehlhaar, D. K.; Verkhivker, G. M.; Rejto, P. A.; Sherman, C. J.; Fogel, D. B.; Fogel, L. J.; Freer, S. T. Molecular recognition of the inhibitor AC-1343 by HIV-l protease: conformationally flexible docking by evolutionary programming. *Chem. Biol.* **1995**, *2*, 317–324.

(47) Verkhivker, G. M.; Rejto, P. A.; Gehlhaar, D. K.; Freer, S. T. Exploring the Energy Landscapes of Molecular Recognition by a Genetic Algorithm: Analysis of the Requirements for Robust Docking of HIV-1 Protease and FKBP-12 Complexes. *Proteins: Struct., Funct., Genet.* **1996**, *250*, 342–353.

(48) Marrone, T. J.; Luty, B. A.; Rose, P. W. Discovering high-affinity ligands from the computationally predicted structures and affinities of small molecules bound to a target: A virtual screening approach. *Perspect. Drug Discovery Des.* **2000**, *20*, 209–230.

(49) Buchdunger, E.; Cioffi, C. L.; Law, N.; Stover, D.; Ohno-Jones, S.; Druker, B. J.; Lydon, N. B. Abl Protein-Tyrosine Kinase Inhibitor STI571 Inhibits In Vitro Signal Transduction Mediated by c-Kit and Platelet-Derived Growth Factor Receptors. *J. Pharmacol. Exp. Ther.* **2000**, *295*, 139–145.

(50) Klabunde, T. Chemogenomics Approaches to Ligand Design. In *Ligand Design for G Protein-coupled Receptors*; Didier, R., Ed.; Wiley-VCH Verlag GmbH & Co. KGaA: 2006; pp 115−135.

CI800138N