Journal of
CHEMICAL INFORMATION
AND MODELING

Article

pubs.acs.org/jcim

Visually Interpretable Models of Kinase Selectivity Related Features Derived from Field-Based Proteochemometrics

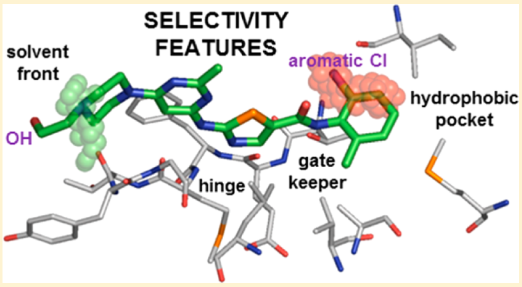
Vigneshwari Subramanian,[†,‡,§] Peteris Prusis,[†] Lars-Olof Pietilä,[†] Henri Xhaard,[‡] and Gerd Wohlfahrt*,[†]

[†]Computer-Aided Drug Design, Orion Pharma, Orionintie 1, FIN-02101 Espoo, Finland

[‡]Centre for Drug Research, Faculty of Pharmacy, University of Helsinki, FIN-00014 Helsinki, Finland

[§]Division of Pharmaceutical Chemistry, Faculty of Pharmacy, University of Helsinki, FIN-00014 Helsinki, Finland

Supporting Information



ABSTRACT: Achieving selectivity for small organic molecules toward biological targets is a main focus of drug discovery but has been proven difficult, for example, for kinases because of the high similarity of their ATP binding pockets. To support the design of more selective inhibitors with fewer side effects or with altered target profiles for improved efficacy, we developed a method combining ligand- and receptor-based information. Conventional QSAR models enable one to study the interactions of multiple ligands toward a single protein target, but in order to understand the interactions between multiple ligands and multiple proteins, we have used proteochemometrics, a multivariate statistics method that aims to combine and correlate both ligand and protein descriptions with affinity to receptors. The superimposed binding sites of 50 unique kinases were described by molecular interaction fields derived from knowledge-based potentials and Schrödinger's WaterMap software. Eighty ligands were described by Mold^2, Open Babel, and Volsurf descriptors. Partial least-squares regression including cross-terms, which describe the selectivity, was used for model building. This combination of methods allows interpretation and easy visualization of the models within the context of ligand binding pockets, which can be translated readily into the design of novel inhibitors.

■ INTRODUCTION

The human genome encodes 518 genes for protein kinases. Kinases are among the most important drug targets, as they play crucial roles in various processes such as cell growth, differentiation, and apoptosis as well as intracellular signal transmission. Not surprisingly, several hundred diseases are related to dysregulation of kinases.[1] Twenty-two kinase-targeting small-molecule inhibitors, including imatinib, lapatinib, dasatinib, gefitinib, and sunitinib, are currently approved mainly for the treatment of cancer.[2] The majority of the kinase inhibitors currently available on the market do interact with multiple targets, frequently leading to toxic side effects.[3,4] Achieving selectivity toward a single biological target has been a main focus of drug discovery programs but has been proven difficult for kinases because of the high similarity among their ATP binding pockets. Several kinase inhibitors have been abandoned because of toxicity, frequently related to low selectivity toward other kinases.[4] Other kinase inhibitors have shown a lack of efficacy, as the use of redundant kinase signaling pathways can limit the effect of (too) specific compounds. Another efficacy-limiting effect is the rise of resistance mutations in the oncogenes. These aspects make it necessary to find the right balance between selectivity and hitting enough targets to overcome immediate resistance.[5] Compounds with targeted polypharmacology offer possibilities to overcome resistance and can show improved efficacy.

An approach widely used in drug design to study the selectivity profiles of compounds is quantitative structure—activity relationship (QSAR) modeling, which involves correlation analysis of the interactions of a series of analogous compounds against a particular protein target. A major drawback of this approach is that it depends only on ligand description and does not include the target—ligand interaction space.[6,7] In order to predict ligand specificity, it is crucial to analyze the target—ligand interaction space, that is, the interactions of multiple ligands across multiple proteins. Proteochemometrics, a multivariate statistical method that aims to correlate both ligand and protein descriptions with affinity to receptors, is well-suited for that task.[7,8] Proteochemometric models provide good predictability and interpretability for both activity and specificity simultaneously for ligands and for targets.[6,7]

Proteochemometrics approaches have been applied to many different protein target families, including G protein-coupled receptors,[8,9] proteases,[10,11] lyases,[12] antibodies,[13] P450s,[14] HLA-DRB1 and MHC protein—peptide interactions,[15,16] and transport proteins[17] as well as kinases.[18,19] Most of these studies used sequence information to describe the target space, although there are a few articles where, for example, local protein substructures[12] or pharmacophoric features of binding

Received: June 27, 2013
Published: October 12, 2013

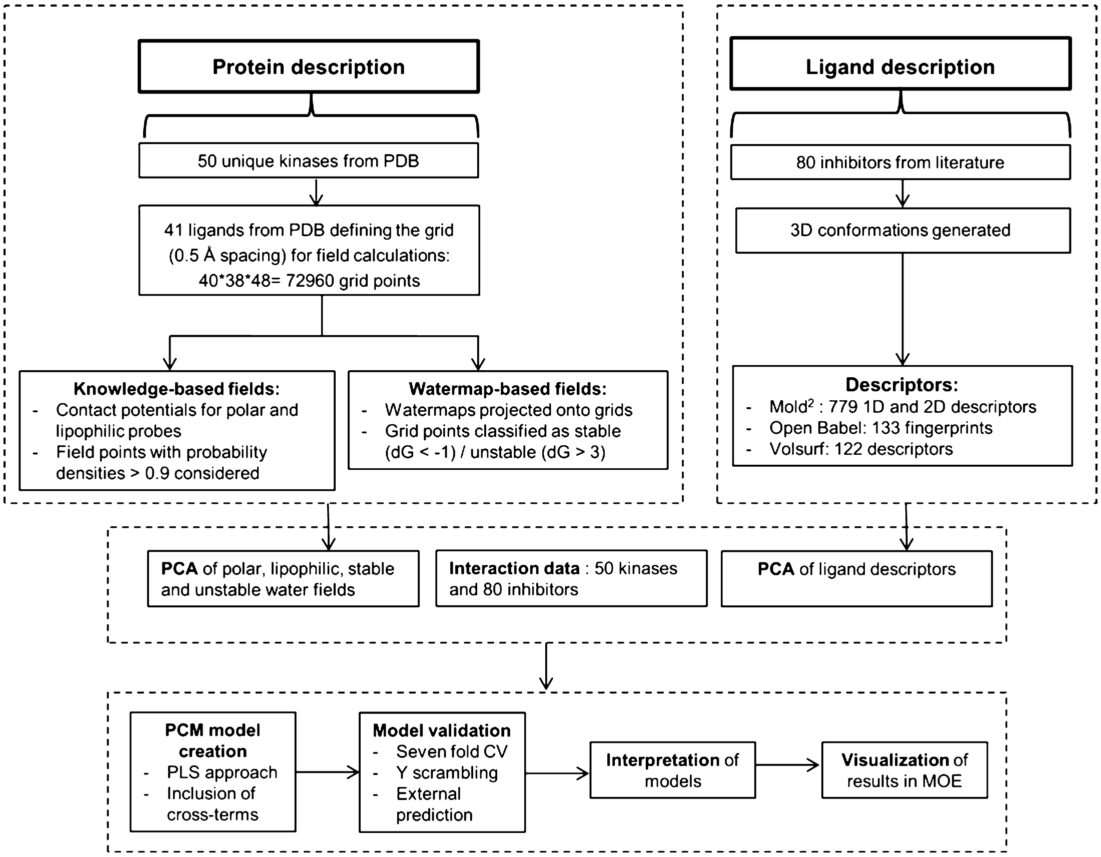

Protein description
50 unique kinases from PDB
41 ligands from PDB defining the grid (0.5 Å spacing) for field calculations: 40*38*48= 72960 grid points
Knowledge-based fields:
- Contact potentials for polar and lipophilic probes
- Field points with probability densities > 0.9 considered
Watermap-based fields:
- Watermaps projected onto grids
- Grid points classified as stable (dG < -1) / unstable (dG > 3)
Ligand description
80 inhibitors from literature
3D conformations generated
Descriptors:
- Mold² : 779 1D and 2D descriptors
- Open Babel: 133 fingerprints
- Volsurf: 122 descriptors
PCA of polar, lipophilic, stable and unstable water fields
Interaction data : 50 kinases and 80 inhibitors
PCA of ligand descriptors
PCM model creation
- PLS approach
- Inclusion of cross-terms
Model validation
- Seven fold CV
- Y scrambling
- External prediction
Interpretation of models
Visualization of results in MOE

Figure 1. Flowchart of the steps involved in the generation and interpretation of field-based proteochemometric models.

sites were used.[20] The comparison of molecular interaction fields of binding sites within a protein family has been shown to be a valuable tool to interpret the selectivity of ligands.[21,22] To our knowledge, the use of molecular interaction fields for target description in proteochemometric modeling has not been applied.

In this article, we demonstrate the creation of proteochemometric models using field-based descriptions of the binding sites of kinases. We have developed and automated several methods that create and compare polar and lipophilic knowledge-based fields[21] or include electrostatic information.[23] The desolvation of binding pockets is a main determinant of ligand affinity and difficult to estimate, for example, with molecular interaction fields. Schrödinger's WaterMap software,[24] which is based on molecular dynamics and statistical thermodynamics, predicts positions and energetics of water in ligand binding pockets. Therefore, in addition to the above-mentioned knowledge-based fields, we also used fields derived from WaterMaps to describe the kinase's ligand binding sites.[25] Employing molecular interaction fields to describe proteins in combination with ligand descriptors in proteochemometric models provides a way to visualize, understand, and modify kinase selectivity profiles of small-molecule inhibitors. The main goal of our study was to create visually interpretable proteochemometric models that support the identification of features relevant for affinity and selectivity in both ligands and receptors.

■ MATERIALS AND METHODS

Software. Schrödinger's nodes (protein preparation and Prime side-chain sampling)[26] were used in KNIME[27] workflows to automate the process of protein preparation. Schrödinger's Prime 3.0[26] was used to introduce missing residues to protein structures. MOE[28] with our in-house SVL scripts was used to perform the knowledge-based-field calculations for the ligand binding sites of kinases. Schrödinger's WaterMap[24] was used to predict stable and unstable water sites in the binding pockets. The manual modeling work and visual inspection was performed using the Schrödinger Maestro interface.[29] The Protein Feature server (PROFEAT) was used to compute the protein descriptors for sequence-based proteochemometric models.[30] MOE 2011.10[28] was used to convert the SMILES notations of the kinase inhibitors into SDF format. Schrödinger's Ligprep 2.5[31] was used to generate 3D structures of ligands; multiple conformations were generated using Confgen 2.3.[32,33] Canvas 1.5[34,35] was used to calculate MACCS fingerprints in order to classify ligands. Mold^2,[36] Open Babel,[37,38] and Volsurf[39] were used to calculate ligand descriptors. SIMCA-P (version 12)[40] was used in proteochemometric modeling. MOE[28] and Pymol[41] were used to make graphical representations of protein–ligand complexes. A workflow summarizing all of the relevant steps during model generation and analysis is shown in Figure 1.

Protein Structures. We collected 144 unique kinase structures from the Protein Data Bank (PDB). If multiple entries were available for a protein, resolution (<3 Å) and completeness were taken into consideration. The different binding modes to active (DFG-in) and inactive (DFG-out) conformations were separated by manual inspection considering the orientation of the DFG motif. There were 122 structures classified as active and 22 as inactive. This study focused only on the active conformation (DFG-in) structures. Of the 122 DFG-in structures, 50 kinases for which activity data

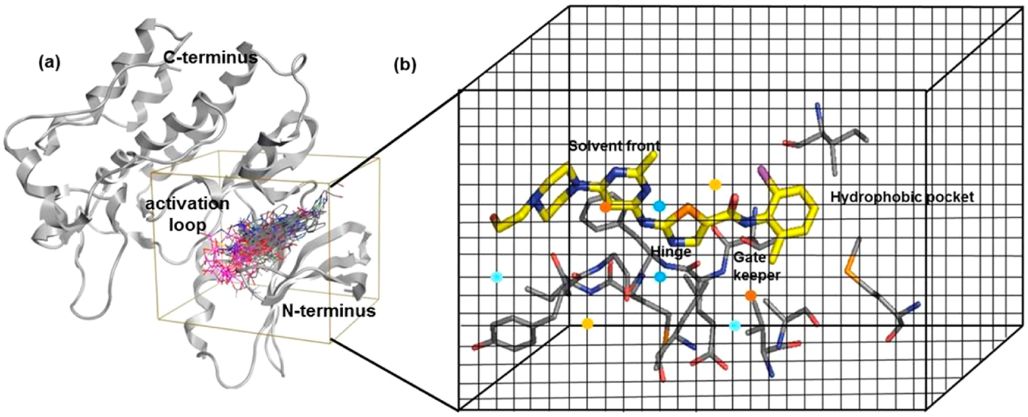


Figure 2. (a) Grid for calculations of knowledge-based fields. The 41 ligands used to define the grid (brown) are shown in the ligand binding site. The structure of ABL1 kinase (PDB ID 2GQG) is shown in gray as a reference. (b) Schematic representation of the grid enclosing the ligand binding site of ABL1. For clarity, only selected residues and a grid with lower resolution are shown. Blue and orange spheres represent polar and lipophilic field points, respectively. Varying color intensities illustrate different probability densities of the field points.

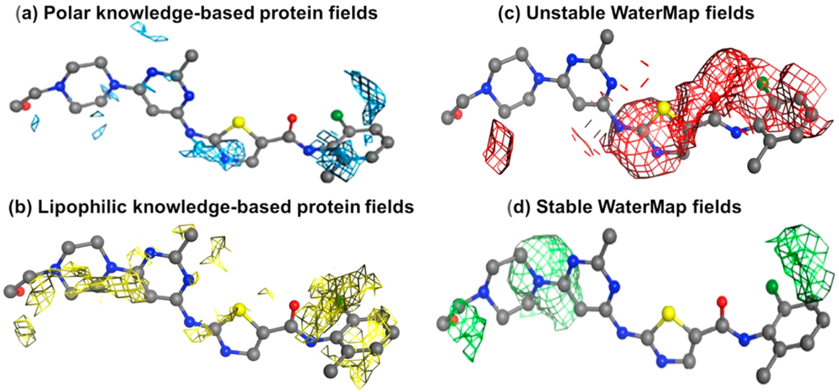

(a) Polar knowledge-based protein fields
(c) Unstable WaterMap fields
(b) Lipophilic knowledge-based protein fields
(d) Stable WaterMap fields

Figure 3. Knowledge-based and WaterMap-based fields calculated from the ligand binding site of ABL1 kinase. The inhibitor dasatinib extracted from the X-ray structure PDB ID 2GQG is shown as a reference. (a) Hydrophilic protein fields with probability densities of >0.9. (b) Lipophilic protein fields with probability densities of >0.9. (c) Unstable water fields with $\Delta G > 3$ kcal/mol. (d) Stable water fields with $\Delta G < -1$ kcal/mol.

were available for at least 80 compounds were chosen for further analysis (see Figure S1 and Table S1 in the Supporting Information). The kinase structures were cleaned by removing additional chains, non-kinase domains, water molecules, and ligands. If the structures had two identical chains, then the more complete chain was used. For five kinases, minor corrections of amino acid residues (modeling residues with missing atoms) were performed. We extracted 41 ligands found in the protein structures so that they could be used as references. Proteins were prepared using a KNIME[27] workflow that performed the following operations: correction of residues with missing atoms, assignment of protonation states, addition of hydrogen atoms, and minimization of hydrogen atoms by keeping heavy atoms fixed. Protein structures with missing residues near the binding site were treated individually and those residues were modeled using PRIME[26] side-chain prediction in Maestro.[29] Following protein preparation, superimposition was performed on a common reference structure (MET, PDB ID 3A4P) using the protein structural alignment tool in Maestro. The average Cα RMSD resulting from superimposition was 2.6 Å, which suggests a successful superimposition procedure, as corroborated by visual analysis.

Ligand Data. We extracted interaction data for 80 compounds binding to 50 kinases from experimental measurements (K_d, K_i) published in three different articles (see Table S2 in the Supporting Information). The selected compounds are likely to bind to the DFG-in conformation according to the literature[42,43] and on the basis of their similarity to known DFG-in inhibitors. We first extracted 33 known DFG-in inhibitors from the data set published by Karaman et al.,[44] which were later used as reference structures to identify further potential DFG-in inhibitors. On the basis of their MACCS fingerprints, we further selected 51 compounds sharing a Tanimoto similarity of >0.8 with the 33 reference compounds. Of the selected 84 compounds, four compounds for which interaction data were available for less than three kinases were removed. The final data set comprised 951 inhibition values for 80 inhibitors measured against a panel of 50 kinases (see the Supporting Information). Of the 2631 observations (protein—ligand combinations) with known binding affinities, only 951 having discrete pK_d/pK_i values were included in PLS models. The remaining 1680 observations were classified as inactives and excluded from further studies. We collected an external test set of 25 compounds that were different from the training set although sharing some similarity; this test set contained 309 activity points (pK_d > 5) for the same 50 kinases.

Description of Kinases. Kinase description was derived from knowledge-based and WaterMap-derived fields of ligand binding sites. Knowledge-based contact potentials, which are expressed as a joint density of three 1D probability densities (interatomic distance, lone-pair interaction angle, and out-of-plane angle), utilize structural information available in the PDB, therefore providing a robust way to describe target−ligand interactions.[28] In order to calculate knowledge-based potentials, first a grid spanning the binding site in the aligned structures was defined, taking advantage of the crystallographic poses of reference ligands. This grid had a spacing of 0.5 Å with limits defined as within 2 Å of all 41 ligands from the X-ray structures (Figure 2a). Then, knowledge-based contact potentials for hydrophilic and hydrophobic probes were calculated for each grid point.[21,23] The probability of each grid point to be hydrophilic or hydrophobic was derived from the distribution of atoms in proximity to that point (see Figure 2b). Only those grid points whose contact probabilities were above 0.9 were considered as relevant. Polar and lipophilic protein field points satisfying the above-mentioned criteria were used as protein descriptors in proteochemometric models (Figure 3a,b).

Ligand affinity is strongly determined by the displacement of water from binding sites. Most water molecules in binding sites are entropically unfavorable, but depending on their hydrogen-bonding interactions they can be enthalpically favorable. Schrödinger's WaterMap,[24] which is based on molecular dynamics and statistical thermodynamics, predicts the position and energetics of water in the binding pocket, resulting in a distinct pattern of stable and unstable water sites.[25] We used fields derived from WaterMaps to describe the kinase's ligand binding sites. WaterMap fields were calculated with the default settings and projected onto the grids used for the knowledge-based methods. Water densities were assigned to each of the grid points. Only grid points whose water density values were above 0.06 were considered relevant and subsequently classified as unstable or stable water depending on their Gibbs free energy values ($\Delta G > 3$ kcal/mol = unstable; $\Delta G < -1$ kcal/mol = stable) (Figure 3c,d). WaterMap-derived fields along with the knowledge-based fields were used as kinase descriptors in proteochemometric models. For comparison with the field-based descriptors, we also used sequence-based properties such as composition, distribution, and order of amino acids to describe kinases.

Description of Kinase Inhibitors. Structures of kinase inhibitors published by Karaman et al.[44] and Davis et al.[42] were downloaded from the PubChem database in SDF format, whereas for those inhibitors published by Metz et al.[45] were generated in SDF format from their SMILES notation. Compounds were characterized by 777 1D and 2D Mold^2 descriptors.[36] Babel's FP4 fingerprints[37,38] (133 fingerprint patterns) calculated on the basis of predefined SMARTS patterns were additionally used to describe the inhibitors. As 3D structural information might provide more relevant information regarding the binding modes of inhibitors, we used Volsurf descriptors to describe the inhibitors by means of field-based numerical description.[39] Prior to Volsurf descriptor calculations, 3D conformations and possible ionization states were generated using the Ligprep module of the Schrödinger suite by applying the default settings. In order to explore the conformational space further, we used Confgen's comprehensive mode to generate multiple conformations for each ligand. However, only the conformation with the lowest potential energy was considered for each ligand and subjected to Volsurf descriptor calculations.

Principal Component Analysis (PCA). PCA is a dimensionality reduction technique that extracts relevant information from large amounts of data by projecting the data onto a lower-dimensional space. After PCA, new attributes called principal components are formed by means of orthogonal transformation of the original attributes.[46] The transformed values constitute the PCA scores and can be defined by

$$\mathbf{T} = \mathbf{XL} \tag{1}$$

where **T** refers to the PCA score matrix of descriptors, **X** is the descriptor matrix, and **L** is the loadings matrix. The high dimensions of the protein descriptors (72 960 field points for each kinase) and ligand descriptors (777 Mold^2, 133 Open Babel, and 122 Volsurf for each ligand) entailed the use of PCA. After all of the variables were scaled to unit variance (see Scaling, Centering, and Variable Transformation below for details), PCAs of polar protein fields, lipophilic protein fields, unstable water fields, stable water fields, amino acid sequence descriptors, and ligand descriptors were performed separately to enable easy interpretation. Including all of the possible principal components can lead to model overfitting. In order to reduce the effects of noise, we extracted only those components whose eigenvalues were above 1. Information concerning the number of components extracted for each descriptor block along with the variation explained is shown in Table 1. The distance of external observation to the PCA model was assessed using DModX as implemented in SIMCA.[40]

Table 1. Principal Component Analysis (PCA) of Protein and Ligand Descriptors

descriptors	components extracted	variation explained (%)
polar protein fields	22	55.0
lipophilic protein fields	23	58.4
unstable water fields	18	64.6
stable water fields	17	67.5
amino acid and dipeptide composition	19	64.7
composition, transition, and distribution	15	78.0
pseudo amino acid composition, sequence order	14	78.3
Mold^2	13	86.0
Open Babel	15	83.6
Volsurf	12	87.7

Cross-Interactions of Ligands and Proteins. In addition to protein and ligand descriptors, nonlinear interactions existing between the proteins and ligands were defined by cross-terms that were expressed as the product of the protein and ligand descriptor variables.[47] It was not computationally feasible to compute the product of the protein and ligand descriptors as such (72 960 protein field points × 133 fingerprints would result in 9 703 680 cross-terms in Open Babel models). For practical reasons, we computed the cross-terms by multiplying the PCA scores for the protein and ligand descriptors.

Scaling, Centering, and Variable Transformation. Before PCA was computed, both protein and ligand descriptors were scaled to unit variance, whereas prior to partial least-squares regression (PLS) the ligand descriptors were mean centered; the cross-terms and protein descriptors were scaled to unit variance. All of the variables were further subjected to block scaling. Each block had a variance of $1/\sqrt{b}$, where b refers to the number of variables in that block.[40] Furthermore, variables of each block were multiplied by the corresponding block weights, which were optimized as described in Table S3 in the Supporting Information. Binding affinity values (K_d or K_i) were subjected to negative $\log_{10}$ transformation, as their distribution appeared to be skewed.

Partial Least Squares Regression (PLS). A PLS approach was used in proteochemometric modeling to identify the correlation between protein/ligand descriptors (X) and binding affinities (Y). The PLS method is similar to PCA, except that components are extracted to achieve maximal covariation between X and Y.[48] Therefore, PLS explains only the linear relationships existing between the variables. We introduced nonlinearity into the models by including cross-terms, which help to identify descriptors that contribute to the selective binding of inhibitors.[6] The proteochemometric models were built using PCA scores of protein and ligand descriptors as the **X** matrix. We created 13 PLS models including different combinations of protein and ligand descriptors. Different weighting parameters were employed to scale the variables in each block, which allowed selection of models with minimal overfit. The model performance was estimated using the squared Pearson's correlation coefficient $R^2(Y)$, which gives the extent of Y variation explained by the model.[40] In addition to the proteochemometric models, we also built for each of the 50 kinases a QSAR (PLS) model based on the PCA scores of ligand descriptors. The performance of the QSAR models was then estimated and compared with that of the proteochemometric models.

Model Validation. In order to examine the effects of overfitting, models were validated using three methods: cross-validation, permutation validation, and an external test set. In cross-validation, the data set was split into several groups. Each group in turn was excluded from the data set, and a new model was built using the remaining observations. The activities of the excluded data were then predicted using this partial model. The performance of the cross-validation was assessed using the squared Pearson's correlation coefficient between predicted and observed activities, denoted as Q^2. The SIMCA default sevenfold cross-validation groups were used in PCM model validations.[40] However, for certain targets (MAPKAPK2, AKT2, and BRAF), activity values were reported for fewer than seven ligands. Therefore, for the corresponding QSAR models, we had to use the "leave one out" (LOO) validation approach. Additionally, we validated the PCM models using a "leave one target out" (LOTO) approach wherein all of the observations for each single target were excluded one by one and the model was built on the remaining targets. The excluded observations were then predicted, and their root-mean-square error of prediction (RMSEP) values computed. The model's ability to fit and cross-validate on random data was probed using permutation validation. Affinity values were reordered 20 times (SIMCA default), and the models were refitted to the permuted data. The performance of permutation validation is characterized by intercepts obtained by plotting the correlation coefficient of the original and permuted values against the R^2 and Q^2 values.[49]

The predictive power of the models was critically assessed by testing them on a separate set of 25 compounds having 309 activity data points. This test set was generated by searching compounds from a pool of 1540 compounds (1527 compounds from the published data sets and 13 compounds from ChEMBL) structurally related to training set compounds. Similarity was assessed by clustering with the training set compounds on the basis of their Open Babel fingerprints using the hierarchical average linkage clustering approach. The model was then used to predict the activities of these 25 compounds, and RMSEP was used to assess the performance.[40]

Model Interpretation. The correlation between the descriptors (X) and binding affinities (Y) can be defined by[50]

$$\mathbf{Y}_c = \mathbf{Y}_m + \mathbf{C}_L\mathbf{X}_L + \mathbf{C}_P\mathbf{X}_P + \mathbf{C}_{LP}(\mathbf{X}_L\mathbf{X}_P) + \mathbf{Y}_e \tag{2}$$

where $\mathbf{Y}_c$ represents the computed Y values; $\mathbf{Y}_m$ corresponds to the mean Y values; $\mathbf{C}_L$, $\mathbf{C}_P$, and $\mathbf{C}_{LP}$ are the coefficients of ligands, proteins, and cross-terms, respectively; $\mathbf{Y}_e$ is the residual; $\mathbf{X}_L$ is the ligand descriptor matrix; $\mathbf{X}_P = [X_{Po1}\cdots X_{Pon}]$ $[X_{F1}\cdots X_{Fn}]$ $[X_{S1}\cdots X_{Sn}]$ $[X_{U1}\cdots X_{Un}]$ is the protein descriptor matrix; and $\mathbf{X}_L\mathbf{X}_P$ is the cross-term. Here, X_{Po}, X_F, X_S, and X_U correspond to the PCA scores of polar protein fields, lipophilic protein fields, stable water fields, and unstable water fields, respectively, and n refers to the number of principal components used in the PLS models. Descriptors that contribute to the binding affinity were evaluated by analyzing the PLS coefficients. For each of the descriptors, the top 10 positive and negative PCA scores were extracted to identify the ligands/proteins whose binding affinities are actually influenced by these components. Features (ligand functional groups/protein field points) related to binding affinity were identified by examining the loadings. Features that support the selective binding of inhibitors can be best understood on the basis of the cross-terms. Because of the large number of cross-terms, we limited the interpretation process by considering only the top five cross-terms that positively influence the binding affinity for each kinase−inhibitor interaction pair.

■ RESULTS AND DISCUSSION

Data Collection. Empirical models are sensitive to the quality of the underlying data because the final equations describing the relationships are not derived from physical laws but rather are adapted to fit the experimental observations. In this work, we therefore selected well-curated sources for training set data, considering three articles[42,44,45] that contain comparable data from well-validated assays. To ensure data quality, we compared the variations in the experimental measurements on the basis of the results published in Fabian et al.[53] and Karaman et al.[44] For a few of the same kinase−inhibitor combinations, considerable differences in experimental values were found (Table S6 in the Supporting Information). This reflects the expected limitations of empirical models imposed by variation in experimental values. Despite the differences even between those rather high quality sources of activity data, for the test set we wanted to assess the model's ability to deal with data of less consistent origin, and therefore, we extracted a set of similar enough compounds directly from the ChEBML database. One of the main goals of our study was to create proteochemometric models that are visually interpretable in the context of ligand binding pockets. Since to our knowledge such efforts have not been reported to date, we wanted to ensure that the molecular properties contributing to the structure−activity relationship (SAR) would be consistent for all of the training set compounds. For several inactives there existed no clear evidence whether they are weak inhibitors or completely inactive, and therefore, we applied an activity cutoff in order to exclude any uncertain cases. The reasons for inactivity in many cases could be completely different from those structure−activity relationships that govern smooth activity and selectivity changes for active ligands. Inclusion of such outlier observations, albeit probably even a minority within the whole set of inactive observations, can

Table 2. Results of Proteochemometric Modeling Using Different Combinations of Ligand Descriptors and Protein Field Descriptors

protein descriptors (field types/sequence)	ligand descriptors	correlation (R^2)	predictivity (Q^2)	RMSEE[a]
polar, lipophilic, watermap[b]	Open Babel	0.336	0.250	0.885
polar, lipophilic, watermap[b]	Mold^2	0.328	0.245	0.892
polar, lipophilic, watermap[b]	Volsurf	0.286	0.217	0.918
polar, lipophilic	Open Babel	0.614	0.437	0.676
watermap	Open Babel	0.599	0.395	0.689
polar, lipophilic, watermap	Open Babel	0.662	0.465	0.633
polar, lipophilic	Mold^2	0.507	0.410	0.764
watermap	Mold^2	0.521	0.421	0.753
polar, lipophilic, watermap	Mold^2	0.539	0.445	0.739
polar, lipophilic	Volsurf	0.489	0.363	0.777
watermap	Volsurf	0.496	0.369	0.772
polar, lipophilic, watermap	Volsurf	0.520	0.400	0.753
amino acid and dipeptide composition	Open Babel	0.517	0.365	0.756
composition, transition, and distribution	Open Babel	0.513	0.343	0.759
pseudo amino acid composition and sequence order	Open Babel	0.484	0.318	0.781

[a] Root-mean-square error of estimation for observations in the training set.[40] [b] Proteochemometric models without cross-terms.

Table 3. Model Validation Based on Cross-Validation, External Prediction, and Permutation Validation

				Y scrambling	
protein descriptors (field types/sequence)	ligand descriptors	RMSEP$_{cv}$[a]	RMSEP[b]	R^2 intercept	Q^2 intercept
polar, lipophilic	Open Babel	0.816	0.823	0.294	−0.125
watermap	Open Babel	0.846	0.836	0.305	−0.115
polar, lipophilic, watermap	Open Babel	0.796	0.80	0.366	−0.107
polar, lipophilic	Mold^2	0.836	0.846	0.240	−0.008
watermap	Mold^2	0.828	0.855	0.250	−0.005
polar, lipophilic, watermap	Mold^2	0.811	0.716	0.278	0.048
polar, lipophilic	Volsurf	0.868	0.942	0.252	−0.005
watermap	Volsurf	0.864	0.958	0.253	0.008
polar, lipophilic, watermap	Volsurf	0.842	0.947	0.293	0.053
amino acid and dipeptide composition	Open Babel	0.867	0.836	0.204	−0.198
composition, transition, and distribution	Open Babel	0.882	0.874	0.179	−0.223
pseudo amino acid composition, sequence order	Open Babel	0.898	0.884	0.175	−0.249

[a] Root-mean-square error of prediction resulting from sevenfold cross-validation. [b] Root-mean-square error of prediction calculated using the external test set.

easily invalidate empirical models and the corresponding interpretation. Applying this reasoning, we considered 951 data points, which in our opinion served as a large enough set of data for the empirical modeling.

QSAR versus Proteochemometrics. In order to compare the performance of the QSAR and proteochemometric modeling approaches, we generated QSAR models for each of the 50 kinase targets (see Table S4 in the Supporting Information). Cross-validation failed for 14 targets, whereas for 14 further targets the models were overfitted. Therefore, it was possible to generate reliable QSAR models for only 44% of the targets in the data set covering less than 50% of the activity values used for training the proteochemometric models. The best QSAR models were obtained for targets with the most uniform and broadest distributions of activity values. Small data sets for some targets together with the inability of ligand descriptors to capture the features important for binding seemed to be responsible for the low Q^2 values and model overfitting, therefore supporting the relevance of generating proteochemometric models involving both protein and ligand descriptors.

Proteochemometric Models. We built proteochemometric models with different combinations of field-based protein descriptors or sequence-based descriptors in order to understand the influence of each of these descriptors on the predictivity. The best R^2 and Q^2 values were obtained by downscaling the protein descriptors and upscaling the ligand descriptors. No significant differences in R^2 and Q^2 were observed when knowledge-based fields or WaterMap-derived fields were used separately (Table 2). However, inclusion of knowledge-based and WaterMap-derived fields together in the same model improved R^2 and Q^2 by 3−5%. The same trend was observed irrespective of the ligand descriptors used. As one can see in Table 2, sequence-based models with Open Babel descriptors have lower R^2 and Q^2 values compared with the corresponding proteochemometric models based on all protein fields. Therefore, we conclude that field-based descriptors possess additional information relevant for activity in comparison with simpler sequence-based descriptors.

As shown in Table 2, proteochemometric models require the inclusion of cross-terms for better performance, which results in a considerable increase in the number of descriptors. Even after using PCA to reduce the number of descriptors, we obtained 1200 cross-terms in Open Babel models, for instance. It is well-known that the inclusion of many descriptors is likely to contribute to spurious random correlations.[49,51] One way to

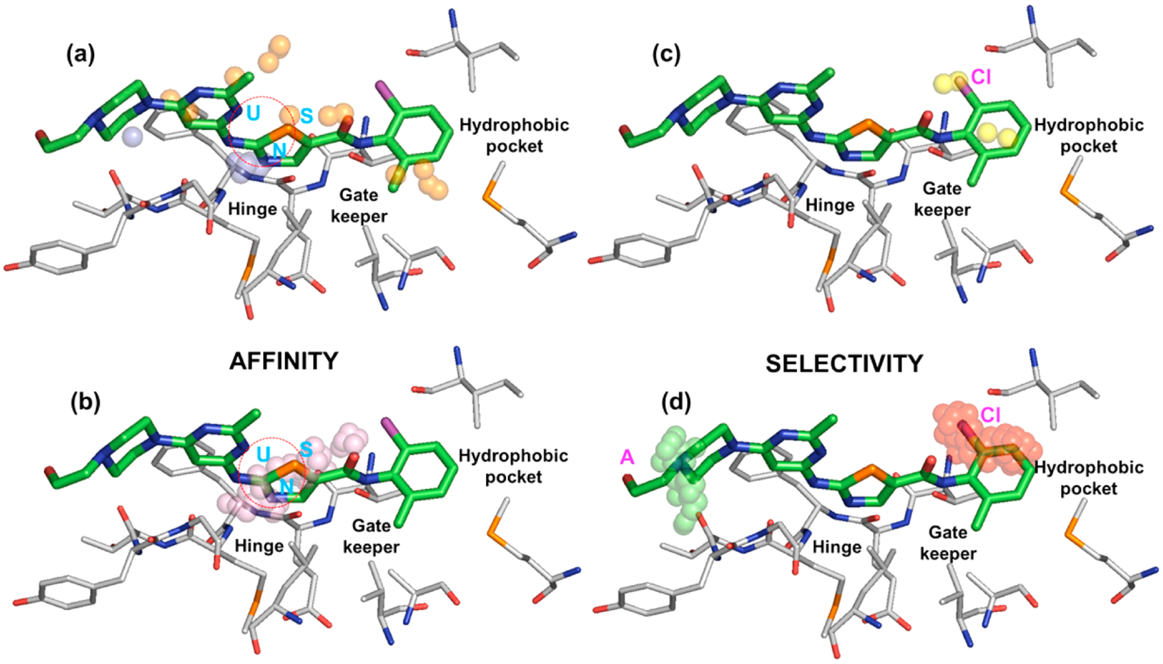


Figure 4. Protein fields and ligand functional groups important for the interactions of ABL1 kinase (gray) with dasatinib (green). Ligand functional groups (from Open Babel) identified as relevant for general kinase affinity (N = hetero-N-nonbasic, S = hetero-S, U = isothiourea) and for selectivity (Cl = aryl chloride, A = primary alcohol) are marked in blue and magenta respectively. (a) Polar and lipophilic protein fields that influence the ligand affinities are represented as slate and orange spheres, respectively. (b) Unstable WaterMap fields that influence the affinity are represented as pink spheres. (c) Lipophilic protein fields that influence the kinase selectivity are represented as yellow spheres. (d) Unstable and stable WaterMap fields that influences the selectivity are represented as red and lime spheres, respectively.

assess the possibility of obtaining high R^2 and Q^2 values by pure chance (i.e., overfitted models) is to use permutation validation.[49] The R^2 intercepts obtained by permutation validation for the models with cross-terms ranged from 0.24 up to 0.37 (Table 3), but the Q^2 intercepts had values close to zero or even negative. It has been stated that a model with R^2 and Q^2 intercepts below 0.3 and 0.05, respectively, is considered to be a valid model.[52] The rather high R^2 intercepts indicate that some of the models with cross-terms might be overfitted. However, the low values of the Q^2 intercepts clearly demonstrate that the reported Q^2 values of the models were not obtained by pure chance. Therefore, we conclude that despite the large number of descriptors, suggesting vulnerability to overfitting, these models are valid enough to be considered for further interpretations.

In order to validate the models further, we predicted the binding affinities of the test set ligands (Figure S2 in the Supporting Information). The RMSEPs of the test set compounds are similar to those obtained from internal validation (see Table 3), except for the Volsurf descriptors, where the external-prediction RMSEPs are higher. Moreover, only 30% of the test set compounds lie within the PCA modeling domain, and the RMSEP for those compounds is even better, namely, 0.73 for Open Babel models. Therefore, from the internal and external validation results we can conclude that the model's ability to predict is limited to the obtained RMSEP values. However, as anticipated, the present models have limited ability to predict inactives. Although this result can be explained by the fact that the inactives lie outside the model's activity range, it also supports our speculation that complete inactivity of some ligands is governed by other reasons than suggested by SAR.

To test the robustness of the models further and to understand the kinase ligand and target space better, we performed LOTO validation. For models based on Open Babel fingerprints (Table S5 in the Supporting Information), the average RMSEP resulting from LOTO validation is 0.820 (standard deviation = 0.022), which is slightly higher than the overall RMSEP of 0.796 from sevenfold validation. For some of the kinases (e.g., KIT), their ligands have a broad and evenly distributed range of activity values, and excluding those contributing to high correlation results in poor predictions. Overall, the RMSEPs based on LOTO and sevenfold validations are rather similar, supporting the conclusion that the models were validated well enough in order to proceed to visualization and interpretation of selectivity features.

Visual Interpretation of Models. The visual interpretation of Mold^2 and Volsurf descriptors is not straightforward; therefore, we focused on the interpretation of models based on Open Babel fingerprints. Features relevant for binding affinity and selectivity were interpreted on the basis of the PCA scores and cross-terms, respectively (see Materials and Methods). The field points and ligand functional groups identified as relevant for affinity or selectivity were visualized with MOE,[28] as our SVL scripts support feature extraction, visualization, and analysis.

The first example uses dasatinib in the binding pocket of ABL1 to illustrate structural elements of ligands and protein areas that are relevant for affinity or selectivity and can also be analyzed in greater detail for a selected set of kinases. The presence of polar points (Figure 4a) and unstable water field points (Figure 4b) near the "hetero-N-nonbasic" group (N) of the ligand were identified to be relevant for the affinity of ABL1 and other kinases toward several ligands. The expression "field points important for affinity" means that these field points are not specific for a kinase family/subgroup. These affinity-promoting regions found in many kinases mainly consist of polar protein regions and hydrogen-bond-forming ligand elements, which contribute to the well-conserved hinge interactions formed by most known kinase ligands. Features that are relevant for selectivity are limited to specific protein−ligand combinations. The presence of aryl chloride in close proximity to lipophilic protein field points (Figure 4c) and unstable water field points close to this (Figure 4d) implies that this functional group makes dasatinib more selective for ABL1 than for several other kinases (Figure 5b). These areas are close to the gatekeeper residue and the hydrophobic back pocket, which differ among kinases and have been frequently exploited in the design of selective inhibitors.[54,55] Additionally, the models suggest that, for example, a hydroxyl group that interacts with a region of stable water sites can contribute to the specificity toward a number of further kinases (Figures 4d and 5b).

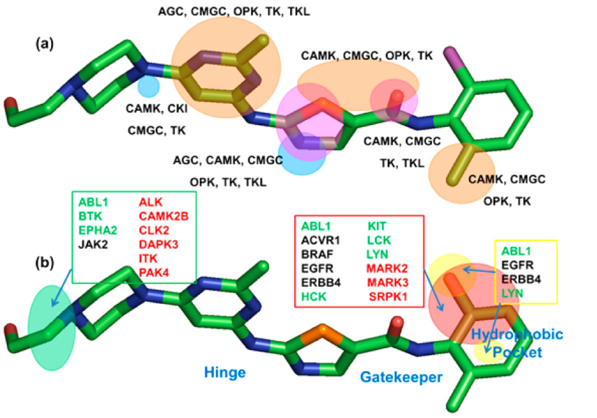


Figure 5. Protein regions that contribute to affinity and selectivity for kinase ligands, mapped to dasatinib in the conformation bound to ABL1. (a) Regions relevant for binding affinity: polar protein fields (blue), lipophilic protein fields (orange), unstable water sites (pink). If a region is relevant for affinity toward a whole kinase group, it is indicated by the group name. (b) Regions relevant for selectivity: lipophilic protein fields (yellow), unstable water sites (red), stable water sites (green). Kinases associated with certain selectivity features and having high (<100 nM), moderate (100 nM to 10 μM), and low/no (>10 μM) binding affinities are listed in green, black, and red, respectively.

Figure 5 illustrates in more detail the structural elements of the protein areas that are relevant for the interactions with either whole kinase groups (affinity) or certain kinases (selectivity). Figure 5a shows the areas that are correlated with affinity to most kinases, annotated here by the main kinase group names. Some protein fields seem less relevant for certain kinase groups. This should not be considered as real selectivity but more as a preference for certain kinase groups. In Figure 5b, areas that should help to distinguish between high and low affinity for certain kinases are shown. This view gives an immediate idea how to modulate the target/antitarget profile of a compound.

In another example with JAK kinase inhibitors, we illustrate the effects of protein and ligand structural elements on potency and selectivity (Figure 6). The affinity of CP-690550A (K_d = 13 nM; Figure 6f) is lower that that of CP-690550 (K_d = 2.7 nM; Figure 6e) as the nitrile has been replaced by a hydroxyl group.[56] The larger nitrile group explores further lipophilic interactions and replaces additional unstable water (Figure 6c,d). While the piperidine C7 methyl (Figure 6a,b) seems to contribute to higher potency for several kinases, the nitrile group probably contributes to the selectivity for JAK kinases.[56] The importance of this area for improved selectivity toward further kinases (Figure 6c,d) is also illustrated by the compound PF-956980, which has an even larger side chain than CP-690550.[57]

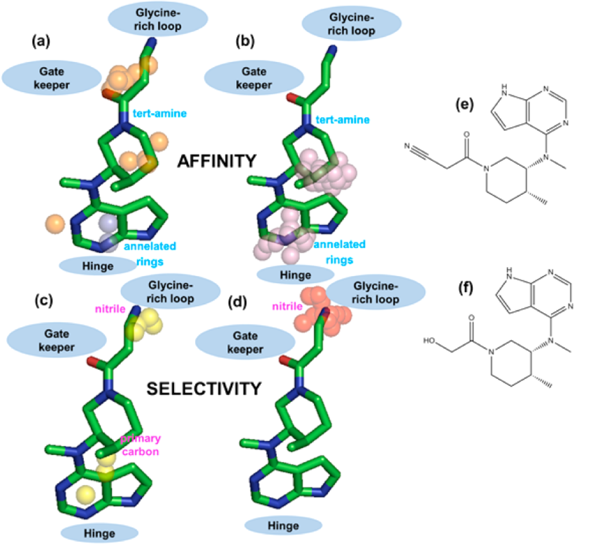


Figure 6. Protein field points and ligand functional groups relevant for the interactions of JAK2 kinases with CP-690550 (shown as a reference). Ligand functional groups (Open Babel) relevant for binding affinity (tertiary amine, annelated rings) and selectivity (nitrile, primary carbon) are marked in magenta and blue respectively. (a) Polar (slate) and lipophilic protein fields (orange) that influence the binding affinity. (b) Unstable WaterMap fields (pink) that influence the affinity. (c) Lipophilic protein fields (yellow) that influence the selectivity. (d) Unstable WaterMap fields (red) that influence the selectivity. (e, f) Structures of (e) CP-690550 and (f) CP-690550A.

CONCLUSIONS

We have shown that the proteochemometrics approach works with protein-derived fields and creates visually interpretable models that can be used to support the design of selective inhibitors. The possibility of investigating the target−ligand interaction space in great detail makes this method combination a promising approach to come up with new ligand design strategies. This is especially the case for ligands targeting larger protein families with well-characterized 3D structures, such as serine proteases, nuclear receptors, and bromodomains. Our present modeling strategy involves manual curation for data preprocessing and interpreting results, and therefore, improved automation of the whole procedure would make it even more suitable for day-to-day inhibitor design work.

ASSOCIATED CONTENT

Ⓢ Supporting Information

Distribution of kinase groups in the data set, external validation plot, molecular weight vs clogP distribution of the training and test set ligands, superimposed kinase structures, experimentally measured affinities used in the studies, training set ligand structures, ligand descriptors (Mold^2, Open Babel, Volsurf), block weights used in the models, QSAR model performance, LOTO validation results, and experimental data showing high variation in the literature. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*Phone: +358-10-4264786. Fax: +358-10-4264682. E-mail: gerd.wohlfahrt@orionpharma.com.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This study was funded by Orion Pharma and by the University of Helsinki. The Finnish IT Center for Science (CSC-IT) is thanked for organizing computational resources at the University of Helsinki, which includes an academic license for Volsurf+. The Finnish National Doctoral Program in Informational and Structural Biology is thanked for organizing graduate studies and providing support to graduate education.

■ REFERENCES

(1) Melnikova, I.; Golden, J. Targeting protein kinases. *Nat. Rev. Drug Discovery* **2004**, *3*, 993−994.

(2) Blue Ridge Institute for Medical Research. http://www.brimr.org/PKI/PKIs.htm (accessed Sept 20, 2013).

(3) Bamborough, P. System-based drug discovery within the human kinome. *Expert Opin. Drug Discovery* **2012**, *7*, 1053−1070.

(4) Scapin, G. Protein kinase inhibition: Different approaches to selective inhibitor design. *Curr. Drug Targets* **2006**, *7*, 1443−1454.

(5) Morphy, R. Selectively nonselective kinase inhibition: Striking the right balance. *J. Med. Chem.* **2010**, *53*, 1413−1437.

(6) Wikberg, J. E. S.; Lapinsh, M.; Prusis, P. Proteochemometrics: A tool for modeling the molecular interaction space. *In Chemogenomics in Drug Discovery: A Medicinal Chemistry Perspective*; Kubinyi, H., Muller, G., Eds.; Wiley-VCH: Weinheim, FRG, 2004; pp 289−309.

(7) Van Westen, G. J. P.; Wegner, J. K.; Ijzerman, A. P.; van Vlijmen, H. W. T.; Bender, A. Proteochemometric modeling as a tool to design selective compounds and for extrapolating to novel targets. *Med. Chem. Commun.* **2011**, *2*, 16−30.

(8) Prusis, P.; Muceniece, R.; Andersson, P.; Post, C.; Lundstedt, T.; Wikberg, J. E. S. PLS modeling of chimeric MS04/MSH−peptide and MC1/MC3−receptor interactions reveals a novel method for the analysis of ligand−receptor interactions. *Biochim. Biophys. Acta* **2001**, *1544*, 350−357.

(9) Lapinsh, M.; Prusis, P.; Lundstedt, T.; Wikberg, J. E. S. Proteochemometrics modeling of the interaction of amine G-protein coupled receptors with a diverse set of ligands. *Mol. Pharmacol.* **2002**, *61*, 1465−1475.

(10) Prusis, P.; Lapins, M.; Yahorava, S.; Petrovska, R.; Niyomrattanakit, P.; Katzenmeier, G.; Wikberg, J. E. S. Proteochemometrics analysis of substrate interactions with dengue virus NS3 proteases. *Bioorg. Med. Chem.* **2008**, *16*, 9369−9377.

(11) Lapins, M.; Eklund, M.; Spjuth, O.; Prusis, P.; Wikberg, J. E. S. Proteochemometric modeling of HIV protease susceptibility. *BMC Bioinf.* **2008**, *9*, 181.

(12) Strömbergsson, H.; Kryshtafovych, A.; Prusis, P.; Fidelis, K..; Wikberg, J. E. S.; Komorowski, J.; Hvidsten, T. R. Generalized modeling of enzyme−ligand interactions using proteochemometrics and local protein substructures. *Proteins* **2006**, *65*, 568−579.

(13) Mandrika, I.; Prusis, P.; Yahorava, S.; Shikhagaie, M.; Wikberg, J. E. Proteochemometric modelling of antibody−antigen interactions using SPOT synthesised peptide arrays. *Protein Eng., Des. Sel.* **2007**, *20*, 301−307.

(14) Kontijevskis, A.; Komorowski, J.; Wikberg, J. E. Generalized proteochemometric model of multiple cytochrome P450 enzymes and their inhibitors. *J. Chem. Inf. Model.* **2008**, *48*, 1840−1850.

(15) Dimitrov, I.; Garnev, P.; Flower, D. R.; Doytchinova, I. Peptide binding to the HLA-DRB1 supertype: A proteochemometrics analysis. *Eur. J. Med. Chem.* **2010**, *45*, 236−243.

(16) Dimitrov, I.; Garnev, P.; Flower, D. R.; Doytchinova, I. EpiTOP—A proteochemometric tool for MHC class II binding prediction. *Bioinformatics* **2010**, *26*, 2066−2068.

(17) De Bruyn, T.; van Westen, G. J.; Ijzerman, A. P.; Stieger, B.; de Witte, P.; Augustijns, P. F.; Annaert, P. P. Structure-based identification of OATP1B1/3 inhibitors. *Mol. Pharmacol.* **2013**, *83*, 1257−1267.

(18) Lapins, M.; Wikberg, J. E. S. Kinome-wide interaction modelling using alignment-based and alignment-independent approaches for kinase description and linear and non-linear data analysis techniques. *BMC Bioinf.* **2010**, *11*, 339.

(19) Fernandez, M.; Ahmad, S.; Sarai, A. Proteochemometric recognition of stable kinase inhibition complexes using topological autocorrelation and support vector machines. *J. Chem. Inf. Model.* **2010**, *50*, 1179−1188.

(20) Meslamani, J.; Rognan, D. Enhancing the accuracy of chemogenomic models with a three-dimensional binding site kernel. *J. Chem. Inf. Model.* **2011**, *51*, 1593−1603.

(21) Hoppe, C.; Steinbeck, C.; Wohlfahrt, G. Classification and comparison of ligand binding sites derived from grid-mapped knowledge-based potentials. *J. Mol. Graphics Modell.* **2006**, *24*, 328−340.

(22) Naumann, T.; Matter, H. Structural classification of protein kinases using 3D molecular interaction field analysis of their ligand binding sites: Target family landscapes. *J. Med. Chem.* **2002**, *45*, 2366−2378.

(23) Wohlfahrt, G.; Sipila, J.; Pietilä, L. O. Field-based comparison of ligand and coactivator binding sites of nuclear receptors. *Biopolymers* **2009**, *91*, 884−894.

(24) *WaterMap*, version 1.4; Schrödinger, LLC: New York, 2012.

(25) Robinson, D. D.; Sherman, W.; Farid, R. Understanding kinase selectivity through energetic analysis of binding site waters. *ChemMedChem* **2010**, *5*, 618−627.

(26) Schrödinger Suite 2011 Protein Preparation Wizard: *Epik*, version 2.2; Schrödinger, LLC: New York, 2011. *Impact*, version 5.7; Schrödinger, LLC: New York, 2011. *Prime*, version 3.0; Schrödinger, LLC: New York, 2011.

(27) Berthold, M. R.; Cebron, N.; Dill, F.; Gabriel, T. R.; Kötter, T.; Meinl, T.; Ohl, P.; Sieb, C.; Thiel, K.; Wiswedel, B. KNIME: The Konstanz Information Miner. In *Data Analysis, Machine Learning and Applications*; Preisach, C., Burkhardt, H., Schmidt-Thieme, L., Decker, R., Eds.; Studies in Classification, Data Analysis, and Knowledge Organization; Springer: Berlin, 2008; pp 319−326.

(28) *Molecular Operating Environment (MOE)*, version 2011.10; Chemical Computing Group Inc.: Montreal, QC, 2011.

(29) *Maestro*, version 9.2; Schrödinger, LLC: New York, 2011.

(30) Li, Z. R.; Lin, H. H.; Han, L. Y.; Jiang, L.; Chen, X.; Chen, Y. Z. PROFEAT: A web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. *Nucleic Acids Res.* **2006**, *34*, W32−W37.

(31) *LigPrep*, version 2.5; Schrödinger, LLC: New York, 2012.

(32) *ConfGen*, version 2.3; Schrödinger, LLC: New York, 2012.

(33) Watts, K. S.; Dalal, P.; Murphy, R. B.; Sherman, W.; Friesner, R. A.; Shelley, J. C. ConfGen: A conformational search method for efficient generation of bioactive conformers. *J. Chem. Inf. Model.* **2010**, *50*, 534−546.

(34) *Canvas*, version 1.5; Schrödinger, LLC: New York, 2012.

(35) Sastry, M.; Lowrie, J. F.; Dixon, S. L.; Sherman, W. Large-scale systematic analysis of 2D fingerprint methods and parameters to improve virtual screening enrichments. *J. Chem. Inf. Model.* **2010**, *50*, 771−784.

(36) Hong, H.; Xie, Q.; Ge, W.; Qian, F.; Fang, H.; Shi, L.; Su, Z.; Perkins, R.; Tong, W. Mold^2, Molecular descriptors from 2D structures for chemoinformatics and toxicoinformatics. *J. Chem. Inf. Model.* **2008**, *48*, 1337−1344.

(37) The Open Babel Package, version 2.3.0. http://OpenBabel.org (accessed Sept 10, 2011).

(38) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An open chemical toolbox. *J. Cheminform.* **2011**, *3*, 33.

(39) Cruciani, G.; Crivori, P.; Carrupt, P. A.; Testa, B. Molecular fields in quantitative structure−permeation relationships: The Volsurf approach. *J. Mol. Struct.: THEOCHEM* **2000**, *503*, 17−30.

(40) *SIMCA-P*, version 12; Umetrics AB: Umeå, Sweden, 2011.

(41) DeLano, W. L. *The PyMOL Molecular Graphics System*, version 0.90; DeLano Scientific LLC: San Carlos, CA, 2003.

(42) Davis, M. I.; Hunt, J. P.; Herrgard, S.; Ciceri, P.; Wodicka, L. M.; Pallares, G.; Hocker, M.; Treiber, D. K.; Zarrinkar, P. P. Comprehensive analysis of kinase inhibitor selectivity. *Nat. Biotechnol.* **2011**, *29*, 1046−1051.

(43) Uitdehaag, J. C. M.; Zaman, G. J. R. A theoretical entropy score as a single value to express inhibitor selectivity. *BMC Bioinf.* **2011**, *12*, 94.

(44) Karaman, M. W.; Herrgard, S.; Treiber, D. K.; Gallant, P.; Atteridge, C. E.; Campbell, B. T.; Chan, K. W.; Ciceri, P.; Davis, M. I.; Edeen, P. T.; Faraoni, R.; Floyd, M.; Hunt, J. P.; Lockhart, D. J.; Milanov, Z. V.; Morrison, M. J.; Pallares, G.; Patel, H. K.; Pritchard, S.; Wodicka, L. M.; Zarrinkar, P. P. A quantitative analysis of kinase inhibitor selectivity. *Nat. Biotechnol.* **2008**, *26*, 127−132.

(45) Metz, J. T.; Johnson, E. F.; Soni, N. B.; Merta, P. J.; Kifle, L.; Hajduk, P. J. Navigating the kinome. *Nat. Chem. Biol.* **2011**, *7*, 200−202.

(46) Wold, S.; Esbensen, K.; Geladi, P. Principal component analysis. *Chemom. Intell. Lab. Syst.* **1987**, *2*, 37−52.

(47) Eriksson, L.; Johansson, E.; Lindgren, F.; Wold, S. GIFI-PLS: Modeling of non-linearities and discontinuities in QSAR. *Quant. Struct.-Act. Relat.* **2000**, *19*, 345−355.

(48) Geladi, P.; Kowalski, B. R. Partial least squares regression: A tutorial. *Anal. Chim. Acta* **1986**, *185*, 1−17.

(49) Eriksson, L.; Johansson, E.; Wold, S. Quantitative structure−activity relationship model validation. In *Quantitative Structure−Activity Relationships in Environmental Sciences*, 7th ed.; Chen, F., Scuurmann, G., Eds.; SETAC: Pensacola, FL, 1997; pp 381−397.

(50) Lapinsh, M.; Prusis, P.; Uhlén, S.; Wikberg, J. E. S. Improved approach for proteochemometrics modeling: Application to organic compound−amine G protein-coupled receptor interactions. *Bioinformatics* **2005**, *21*, 4289−4296.

(51) Topliss, J. G.; Costello, R. J. Chance correlations in structure−activity studies using multiple regression analysis. *J. Med. Chem.* **1972**, *15*, 1066−1068.

(52) Eriksson, L.; Johansson, E.; Kettaneh-Wold, N.; Wold, S. *Introduction to multi- and megavariate data analysis using projection methods (PCA & PLS)*; Umetrics AB: Umeå, Sweden, 1999.

(53) Fabian, M. A.; Biggs, W. H., III; Treiber, D. K.; Atteridge, C. E.; Azimioara, M. D.; Benedetii, M. G.; Carter, C. A.; Ciceri, P.; Edeen, P. T.; Floyd, M.; Galvin, M.; Gerlach, J. L.; Grotzfeld, R. M.; Herrgard, S.; Insko, D. E.; Insko, M. A.; Lai, A. G.; Leilas, J. M.; Mehta, S. A.; Milanov, Z. V.; Velasco, A. M.; Wodicka, L. M.; Patel, H. K.; Zaarinkar, P. P.; Lockhart, D. J. A small molecule−kinase interaction map for clinical kinase inhibitors. *Nat. Biotechnol.* **2005**, *23*, 329−336.

(54) Das, J.; Chen, P.; Norris, D.; Padmanabha, R.; Lin, J.; Moquin, R. V.; Shen, Z.; Cook, L. S.; Doweyko, A. M.; Pitt, S.; Pang, S.; Shen, D. R.; Fang, Q.; de Fex, H. F.; Mclntyre, K. W.; Shuster, D. J.; Gillooly, K. M.; Behnia, K.; Schieven, G. L.; Wityak, J.; Barrish, J. C. 2-Aminothiazole as a novel kinase inhibitor template. Structure−activity relationship studies toward the discovery of *N*-(2-chloro-6-methylphenyl)-2-[[6-[4-(2-hydroxyethyl)-1-piperazinyl)]-2-methyl-4-pyrimidinyl]amino)]-1,3-thiazole-5-carboxamide (dasatinib, BMS-354825) as a potent *pan*-Src kinase inhibitor. *J. Med. Chem.* **2006**, *49*, 6819−6832.

(55) Zuccotto, F.; Ardini, E.; Casale, E.; Angiolini, M. Through the "gatekeeper door": Exploiting the active kinase conformation. *J. Med. Chem.* **2010**, *53*, 2681−2694.

(56) Chrencik, J. E.; Patny, A.; Leung, I. K.; Korniski, B.; Emmons, T. L.; Hall, T.; Weinberg, R. A.; Gormley, J. A.; Williams, J. M.; Day, J. E.; Hirsch, J. L.; Kiefer, J. R.; Leone, J. W.; Fischer, H. D.; Sommers, C. D.; Huang, H. C.; Jacobsen, E. J.; Tenbrink, R. E.; Tomasselli, A. G.; Benson, T. E. Structural and thermodynamic characterization of the TYK2 and JAK3 kinase domains in complex with CP-690550 and CMP-6. *J. Mol. Biol.* **2010**, *400*, 413−433.

(57) Kudlacz, E.; Conklyn, M.; Andresen, C.; Whitney-Pickett, C.; Changelian, P. The JAK-3 inhibitor CP-690550 is a potent anti-inflammatory agent in a murine model of pulmonary eosinophilia. *Eur. J. Pharmacol.* **2008**, *582*, 154−161.