AutoShim: Empirically Corrected Scoring Functions for Quantitative Docking with a Crystal Structure and IC_{50} Training Data

Eric J. Martin* and David C. Sullivan

Department of Computer Aided Drug Discovery, Global Discovery Chemistry, Novartis Institutes for Biomedical Research, 4560 Horton Street, Emeryville, California 94608

Received December 7, 2007

It has been notoriously difficult to develop general all-purpose scoring functions for high-throughput docking that correlate with measured binding affinity. As a practical alternative, AutoShim uses the program Magnet to add point-pharmacophore like "shims" to the binding site of each protein target. The pharmacophore shims are weighted by partial least-squares (PLS) regression, adjusting the all-purpose scoring function to reproduce IC_{50} data, much as the shims in an NMR magnet are weighted to optimize the field for a better spectrum. This dramatically improves the affinity predictions on 25% of the compounds held out at random. An iterative procedure chooses the best pose during the process of shim parametrization. This method reproducibly converges to a consistent solution, regardless of starting pose, in just 2−4 iterations, so these robust models do not overtrain. Sets of complex multifeature shims, generated by a recursive partitioning method, give the best activity predictions, but these are difficult to interpret when designing new compounds. Sets of simpler single-point pharmacophores still predict affinity reasonably well and clearly indicate the molecular interactions producing effective binding. The pharmacophore requirements are very reproducible, irrespective of the compound sets used for parametrization, lending confidence to the predictions and interpretations. The automated procedure does require a training set of experimental compounds but otherwise adds little extra time over conventional docking.

INTRODUCTION

Limitations of All-Purpose Scoring Functions for Predicting Activity. Although docking is one of the most commonly used methods in computational chemistry, all-purpose scoring functions generally cannot even qualitatively predict experimental binding affinity. In a recent publication[1] of a huge docking survey at GlaxoSmithKline by 15 authors, consisting of experts in both the targets and the docking software, the best correlation they found with activity (−log affinity) was $R^2 = 0.32$. While the study did sometimes find useful pose predictions, and positive enrichment in yes/no activity prediction curves, the conclusion was that "there is no statistically significant correlation between measured affinity and any of the [37] scoring functions [from 10 docking programs] evaluated across all eight protein targets examined."[1]

Alternative of Target-Specific Scoring Functions. What can be done to make docking more predictive? If theory is inadequate, are there other available data that could be empirically employed to improve the predictions? The most commonly available and relevant data in drug discovery projects are IC_{50} activity data. If general all-purpose scoring functions do not work, one possible alternative is to use the available IC_{50} data to parametrize custom scoring functions for each specific target. This paper presents AutoShim, an empirically parametrized, target specific scoring function created by adding pharmacophore-like interaction terms to any all-purpose scoring function. By analogy to NMR, where more or less energy is added to suitably placed magnetic "shims" to create a more homogeneous field, here pharmacophoric shims are added to the general purpose scoring function, augmenting some interactions and attenuating others, to reproduce IC_{50} data better.

This does represent a tradeoff. All-purpose scoring functions do not require initial data for parametrization, and can be theoretically more satisfying, particularly if based on first principles. AutoShim must, therefore, routinely do better than all-purpose scoring functions. The very best docking performance from the nearly 300 attempts in the GSK study, $R^2 = 0.32$, is taken as the criterion for success. Routinely exceeding this bar will justify resorting to a parametrized, target-specific scoring function.

Other target-specific methodologies have been presented in the past. One of the earliest, and most conservative target-specific binding energy estimation methods is Aqvist's "linear interaction energy" (LIE) approach,[2] in which the sum of electrostatic contributions are weighted independently from the sum of van der Waals components and calibrated to experimental data. LIE concentrated on using affinity data with high quality crystal structures for a series of related compounds. Rebecca Wade's COMBINE[3] customizes the molecular mechanics force field in the active site. PLS (see the Abbreviations section for definitions of abbreviations used in this paper) for regression against percent-inhibition data adjusts the weights of individual molecular mechanics force field terms that have been fragmented into thousands of terms reflecting protein residue, ligand functional group, and interaction type. A variable selection procedure reduces the number of descriptors contributing to the final PLS

* To whom correspondence should be addressed. Phone: (510) 923-3306. Fax: (510) 923-2010. E-mail: eric.martin@novartis.com.

10.1021/ci7004548 CCC: $40.75
Published on Web 04/02/2008

equation to around 50. Pharmacophore features associated with docked poses across grid-based volumes have also served as linear regression descriptors. [4] The residue-based SIFt[5] methodology of Singh et al. calculates an interaction "fingerprint" by encoding seven types of yes/no interactions between the ligand and each active site residue (e.g., Does this residue contact ligand? Does a main chain atom contact ligand? polar interaction?). The authors utilize this fingerprint to cluster dockings and find interaction commonalities among crystallized kinase inhibitors. Commercial programs that facilitate calculating interaction descriptors for docking results include the database docking environment of FlexX[6] and the "Silver" postprocessing component of GOLD.[7] These and other target-biased approaches to structure-based virtual screening have been reviewed previously.[8]

The AutoShim method presented in this paper extends the Magnet[9] "expert system" technology for target-specific scoring functions. The Magnet approach started 10 years ago, from frustration over the need to visually review the results of every docking run, and the impossibility of reviewing more than the top few thousand poses. An expert system was envisioned that could be interactively trained to reproduce an expert's reasoning for accepting or rejecting various dockings for each target. Dr. Hanneke Jansen wrote the first prototype, by ingeniously using custom-defined Sybyl[10] spreadsheet columns to define the specific protein–ligand interactions. It allows the user interactively to define ligand binding features, such as hydrogen bonds with specified protein atoms or occupancy of active site pockets by specific types of ligand atoms. These features could be used as filters, or in linear combinations to create a target-specific scoring function. Jansen et al. continued to develop the technology for several years, eventually partnering it with Metaphorics, where Dr. Scott Dixon created a special programming language called "SEA" for very flexibly defining binding interactions. Magnet is now distributed with DockIt by Daylight Chemical Information Systems.

Magnet has been effectively applied as an "art-based" expert system to many problems over many years: filtering hit lists, improving docking results, and clustering docking modes. It might best be described as "high throughput intuition" (HTI). The expert user trains the program to reproduce the decisions of an individual expert or team, on hundreds of examples. The program can then evaluate hundreds of millions of poses as if the user had visually evaluated each one. The ability to capture and encode expert intuition is unique and very useful. Performed interactively as an expert system, however, using Magnet is a slow, manual process. "AutoShim" uses indirect programming to write SEA routines to automatically generate a standard set of simplified point pharmacophore-like Magnet interaction features. It then regresses these Magnet generated features (or multipoint pharmacophores built up from Magnet features) along with docking score and contact terms, against experimental activity data by partial least-squares regression (PLS).[11] Particular protocols for automatically generating features are explored in this paper.

Multiple Instance Problem. Another distinguishing characteristic of AutoShim is its use of iterative pose selection to solve docking's "multiple-instance problem". Up to hundreds of docked poses can be generated for each compound. Yet each compound has only 1 experimental affinity corresponding to a single "best" pose. Ideally, the selected pose would correspond to the crystallographically observed geometry. An ideal all-purpose scoring function would identify the best pose as having the lowest energy, but actual all-purpose scoring functions do not reliably select this pose, even for the "easy" case of redocking a ligand extracted from a holocrystal structure, which should be preformed to bias the correct ligand pose.[12,13] To improve binding mode prediction, scoring functions can be tuned to select the crystallographic pose,[14–16] however we know of no study demonstrating that tuning a general-purpose scoring function for binding mode prediction improves affinity prediction for a diverse set of compounds. It has been shown that selecting a chemotype's binding mode for which structure and docking-predicted activity show the strongest relationships improves the chance of selecting the experimental binding mode.[17] Also, a scoring function trained to select experimental binding poses has been shown to perform well at activity prediction,[18] providing at least circumstantial evidence that accurate binding mode prediction is linked to accurate activity prediction. In the very large GSK study mentioned above, however, Warren et al. observed that "high fidelity in the reproduction of observed binding modes did not automatically impart success in virtual screening. However, of particular concern was the observation that some scoring functions required no correct structural information for success in virtual screening [enrichment curves]."

Regardless of whether the correct pose is needed for accurate affinity prediction, the choice of a pose determines which interactions each compound can make, and therefore is needed to parametrize an empirical scoring function in the first place. This creates a dilemma, however, since the selected pose should be the lowest energy choice by that very scoring function for subsequent predictions. Thus, parametrization and pose-selection must somehow be performed simultaneously. This is known as a "multiple instance problem".

One solution to the multiple instance problem that should be avoided is to use a numerical optimization to choose the pose for each compound that results in the best regression model to correlate predicted affinity with measured affinity. There are at least two problems with this approach. First, this pose selection procedure itself requires experimental data, and thus cannot be applied to predictions of new compounds. Second, the combinatorics of choosing any pose for each training compound gives an astronomical number of combinations that is virtually certain to result in overfitting. This problem is similar to the often-cited overfitting mistake of optimizing CoMFA alignments to produce the best Q^2.

This study solves the multiple-instance problem iteratively by parametrizing the empirical scoring function on the single "current best" pose for each compound. Initially the best pose is from an all-purpose scoring function. PLS then parametrizes an AutoShim scoring function to better reproduce experimental IC_{50} data. After each round of parametrization, the single highest-predicted-activity pose for each compound is selected with the updated model, to train a next-iteration scoring function. In practice, PLS models converge in both performance and descriptor weightings in the PLS equation after just 2–4 iterations, independent of the initial pose set selection. For predictions on new compounds, AutoShim uses the single best scoring pose for each compound as usual.

METHODS

Activity Data Sets. Two principal IC_{50} data sets showcase AutoShim, each drawing on diverse compounds from several medicinal chemistry series as well as diverse purchased compounds. The CSF1R activity data set contains 2149 compounds, of which 294 tested sub-100 nM. The PDK1 data set contains 1913 compounds, with 253 sub-100 nM hits. Of these, 185 compounds are common to the two data sets.

A wide chemical coverage and diversity of these data sets is indicated by counting unique "Bemis and Murcko" molecular frameworks,[19] in which side chains have been stripped off leaving only ring systems and linkers. Disregarding atom types, hybridization, and bond order, the 2149 CSF1R compounds represent 891 molecular frameworks. Even the 294 hits, which one might expect to be more chemically focused, still cover 92 frameworks. The 1913 PDK1-tested compounds cover 697 frameworks, with the hits alone covering 58 frameworks. For comparison, Bemis and Murcko found all of "drug space", as defined by 5120 commercially available drugs in the Comprehensive Medicinal Chemistry database, to be composed of 1179 frameworks, similar in magnitude to our data sets.

Docking. Docking and scoring employ a three step process whereby DockIt[20] generates an initial set of several hundred docked ligand poses per compound, which are filtered by interaction patterns using Magnet,[9] and finally minimized and scored with the Flo+ component of QXP.[21] At this point, poses exhibiting severe steric clashes with the receptor, as well as poses that lose prescribed interactions during minimization, are removed. This workflow captures DockIt's strength in sampling diverse poses stemming from a distance-geometry pose generator, while Magnet filtering reduces the number of poses to minimize. For the kinase family, Magnet is used to filter away poses not making at least one hydrogen bond to the hinge, as known kinase inhibitors generally adopt one of the three hydrogen bonds which stabilize the adenine ring of bound ATP.[22,23]

Magnet. In addition to filtering docking poses, Magnet calculates the various types of molecular interaction features used to "shim" the scoring functions: hydrogen bonding indicators, pocket occupancy indicators and atom-type counts, and counts of ligand atoms contacting protein. Rules for generating Magnet features are coded in the SEA language. A program called "c2sea" was written in C to automatically generate SEA code for point-pharmacophore features from a list of feature coordinates and radii (see below).

PLS. Partial least-squares (PLS) regression[11] was performed in the *R* statistical computing environment[24] using the function "mvr" from the pls package.[25,26] PLS constrains correlated descriptor coefficients to move together, by reducing the independent variables to a smaller set of latent variables. PLS latent variables are linear combinations of the original independent variables that maximize variance, similar to principal components,[27] while also explaining the dependent variable. 5-fold cross-validation determines the optimal number of latent variables. Prediction performance is evaluated on a test set withheld from PLS, as well as from (where applied) data-driven complex descriptor generation (see below).

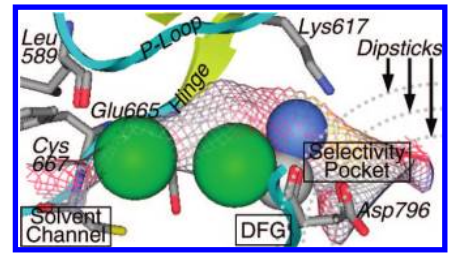

Figure 1. CSF1R Art shims include hydrogen bonds to Leu589 carboxyl, the hinge backbone (Glu665 carboxyl, Cys667 NH, and carboxyl), Lys617 amine, and Asp796 backbone NH. Atoms are counted within spheres of 3 types: aromatic (green), hydrogen bond donor (blue), and any non-hydrogen (grey). Binary dipsticks are set from 0 to 1 by any nonhydrogen atom. The wire mesh indicates CSF1R's interaction surface.

Hand-Built Shims for CSF1R: Art and Art-X. The first empirically parametrized scoring function, specific for CSF1R, used Magnet as an expert system to define art-based protein–ligand interactions identified by the project team as important for activity. These interactions were encoded in the SEA language, extracted for each pose with Magnet, and weights for each interaction along with Flo+ score, were determined by iterative PLS/pose selection against pIC_{50} activity data (see below). Here, 2149 compounds with IC_{50} data were available for modeling, of which a random 25% were held back as an external test set. These hand-built features (Figure 1) stressed parsimony to aid interpretation. Several general categories of Magnet rules (e.g., hydrogen bond and pocket occupancy descriptors) were evaluated and selected by inspecting their contribution to regression equations on a training set of compounds. The "Art" set of features selected to shim the Flo+ score is defined as follows.

(i) Hydrogen bonds are defined by having a specific protein heavy atom lying within 3.3–4 Å of (fixed for any single descriptor rule), but at least 2.3 Å from, any complementary hydrogen bonding ligand atom. Also, the two heavy atom hydrogen bond donors (HBD) and acceptors (HBA) must form angles with all covalently neighboring atoms (both on the protein and ligand) of greater than 80°. Only two protein HBAs contribute to the final hand-built CSF1R descriptor set, one to the Cys667 carbonyl in the hinge and another to the Leu589 carbonyl in the glycine rich loop (Figure 1). Several other backbone and side chain HBAs were also considered during model development, but were not significant. Protein HBDs in the final model include the backbone amides of Cys667 in the hinge, Asp796 in the DFG motif, and to the catalytic amine of Lys617. The total number of hydrogen bonds to the hinge backbone (Glu665 carbonyl, Cys667 carbonyl and amide) defines an additional descriptor, as well as another binary descriptor that is set "on" if all three hinge hydrogen bonds are made.

(ii) Pocket occupancy descriptors are defined by counts of particular atom types within a radius of a feature center. Feature centers were suggested by atom positions in crystallized ligand examples. Atom types employed include aromatic atoms, HBDs, and nonhydrogen atoms. Radii are either 1.5 or 2 Å.

(iii) Whole-molecule and whole-interaction descriptors utilized (in addition to the Flo+ score) include the count of ligand heavy atoms, the number of ligand atoms contacting protein, the number of protein atoms contacting ligand, the ratio of the number of protein atoms making ligand contacts

to ligand atoms making protein contacts, the ligand atom count divided by the number of receptor atoms making ligand contacts, and the steric clash component of the Flo+ score.

(iv) A series of three binary pocket occupancy descriptors using features located in the rear cleft with large radii, define "dipstick" descriptors that assess how deep a ligand binds into the receptor.

(v) A binary descriptor identifies whether a pose comes from the "open" crystal structure or "closed" crystal structure (see below).

These 20 descriptors are used directly as the shims in the Art PLS regression. However, they contain no cross-terms. A second shim set, named "Art-X", added selective cross-terms to the Art set. Inspecting recursive partitioning decision trees of activity using these linear descriptors suggested the feature combination of binding deeply (assessed by one of the dipstick descriptors) AND forming the Leu589 carboxyl hydrogen bond. Finally, all 21 descriptors were crossed (i.e., multiplied) with the Flo+ score and with one of the dipstick descriptors, thus tripling the total number of Art-X shims to 63.

Automated Generation of Interaction Features. Testing and refining art-based shims in the Magnet expert system involves a lot of manual work. A more automated method of generating interaction features was sought, to facilitate routine AutoShim modeling. With an eye toward applying a single descriptor set to docked poses originating from different superimposed kinases,[28] all protein residue-based descriptors were abandoned, instead employing only spatially located pharmacophore-spheres to describe protein–ligand interactions. This strategy required selecting feature positions throughout the binding site. Two possible solutions were considered. First, the protein receptor could be interrogated to find locations with high interaction potential using a program such as GRID[29] or sphgen.[30] Instead we clustered ligand atom positions from receptor-bound crystal costructures (three for CSF1R, eight for PDK1) down to 66 feature centers. This unbiased approach lets the training of the PLS scoring function discover the important interactions, yet focuses descriptor sampling on relevant regions of the active site. Binary descriptors indicate the presence or absence of four ligand atom types (HBD, HBA, aromatic, and hydrophobic) at each feature position. For CSF1R, each feature's radius is individually set at the distance to the nearest neighboring feature center. For PDK1, a fixed radius of 1.5 Å was imposed for all features. The 264 pharmacophore features, along with the ligand atom count, number of ligand atoms contacting protein, and the number of protein atoms contacting the ligand, define the "Auto" descriptor set for shimming Flo+ score via PLS regression, to create the "shimmed" scoring function. Thus, while each pharmacophore feature measures the binary presence/absence of an interaction, the contribution (as a shim) to the docking score is weighted by a continuous coefficient in the final scoring function. These automated descriptors were defined for Magnet using SEA programs created with c2sea.

Recursive Partition Multipharmacophore Autoshims. The automated shim set provides single-point shims (SPS) for PLS, but lacks nonlinear feature combinations, i.e. PLS cross-terms. Employing all pairwise cross-terms would exceed computational limitations and flood the equation with irrelevant descriptors. Instead, recursive partitioning[31] (RP) was recruited as a data-driven method to generate relevant feature combinations for PLS (PLS-on-RP). The RP trees generate binary "rules", reminiscent of multipoint pharmacophores, that serve as multipoint shims (MPS) for subsequent PLS-regression.

This implementation of PLS-on-RP assumes iterative pose selection, which is detailed below. The first 6 of 10 PLS/pose selection iterations in PLS-on-RP include shim building steps as well as the PLS regression itself. At each iteration, six separate recursive partitioning classification trees are trained, each using a different binary activity threshold: 1%, 5%, 10%, 30%, 50%, and 70%. Since fixed percentages of compounds are designated active, the absolute pIC_{50} thresholds differ across data sets. Classification trees are built using the *R* function "rpart" from the rpart package,[31,32] with default parameters except that depth is limited to 7 splits. The tree "leaves" that predict active compounds (active class purity >50%) contribute binary cross-terms (i.e. MPSs) for subsequent PLS regression. Rules for inactivity were rejected, as their inclusion does not improve predictive accuracy, probably because inactivity is often a nonspecific event. In addition, each of the component individual pharmacophore features contributing to the leaves from a seventh tree, with an absolute pIC_{50} activity threshold of 5.8, are included as linear terms (i.e. SPSs) for PLS.

PLS-on-RP takes advantage of the multiple PLS/pose-selection iterations to accumulate shims. For the second iteration of PLS/pose selection, the shims from the first iteration are *replaced* with a new RP training on the features set by the new best poses. However, for the third through sixth iterations, the shims from each newly trained RP tree are *added* to the existing set. The seventh tree, contributing exclusively single-term shims, is rebuilt at each of the first six iterations, discarding the previous iteration's shims. For the final four iterations, the shim set is held fixed, and the PLS model evolves only by pose selection.

Note that the interaction features defined by c2sea, and extracted from each pose by Magnet, are unsupervised shims, independent of the IC_{50} data. The activity-supervised RP converts these unsupervised features into a target-optimized shim set. For the binary descriptors, the RP steps amount to variable and cross-term selection for the subsequent PLS. For continuous and multivalue input descriptors (see accompanying manuscript[28]), recursive partitioning also provides the optimal thresholds for conversion from continuous to binary descriptors. Besides contributing to the recursive partitioning, the Flo+ score is also added as a continuous variable in the PLS descriptor matrix, along with various contact terms described above. The resulting PLS regression equation is the target-customized scoring function.

Because RP is performed outside PLS cross-validation, this could bias toward overtraining with cross-validation selecting too many latent variables. As an ad hoc measure to limit overtraining in PLS, the predictive correlation calculated during cross-validation, Q^2, is penalized by 0.002 multiplied by the number of latent variables prior to assessing the optimal number of latent variables.

Iterative Pose Selection. *Current-Best Low-Energy Pose Selection.* PLS models are initially trained on the Magnet interactions found in each compound's highest Flo+ scoring pose. For each compound in the training set, the pose with the maximum predicted activity (MPA) in the current PLS

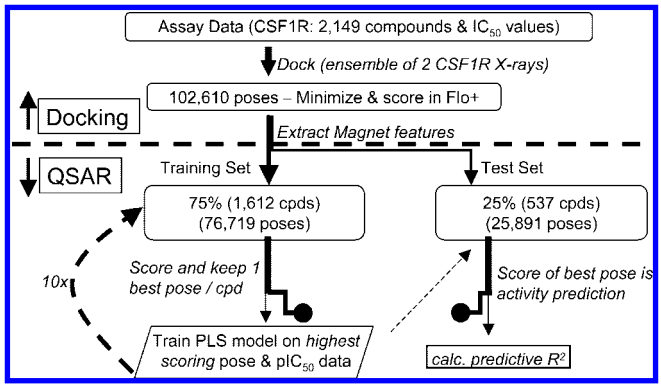

Figure 2. Flowchart of the model building and affinity prediction process for CSF1R.

model is then used in the next training iteration. It is worth emphasizing that the selected pose has the highest score, not the score most closely matching experimental activity. Ten iterations are performed, although convergence generally occurs within just 2−4. This fast convergence, and performance on the 25% withheld test set, demonstrates that overfitting is avoided. The PLS model predicts activities for all test set poses, with the highest prediction per compound defining that compound's activity prediction and pose.

Least Prediction Residual Pose Selection. For comparison to MPA pose selection, we also studied an ill-advised least-prediction-residual (LPR) pose selection. At each iteration, the PLS model predicts activities for all training set poses, and the poses with a prediction closest to the experimental pIC_{50} creates the next iteration's training set (see the Results section).

Ensemble of Proteins Treated Like More Poses. For CSF1R, two predominant active-site conformations and corresponding binding modes were observed among available crystal structures. One family binds deeply into the back cleft, extending past the gatekeeper, which is open in these crystal structures. Another family does not bind so deeply, instead focusing interactions in the ribose pocket located toward the mouth of the active site. These ribose pocket oriented compounds are not able to assume their crystallographic conformations in the open crystallographic conformations. Conversely, crystal structures with ribose-pocket binders display a closed pocket conformation in the vicinity of the gatekeeper, precluding the deep binding found in the open-pocket examples. In order to capture protein flexibility implicitly, dockings against both an open and closed crystal structure create the pool of poses available for AutoShim. A "pose", therefore, includes a protein conformation as well as the usual ligand conformation and orientation.

The AutoShim components, including the *R* code for statistical modeling, c2sea for generating Magnet compatible SEA rules, and accessory scripts, have been deposited with SourceForge.net and daylight.com. The code is available under an open source copyright agreement.

RESULTS

CSF1R. Figure 2 shows a flowchart of the entire procedure for AutoShim model building and predictions for CSF1R. In this procedure, 2149 compounds with IC_{50} data were docked into an ensemble of two X-ray crystal structures yielding 102 610 poses. These were minimized and scored

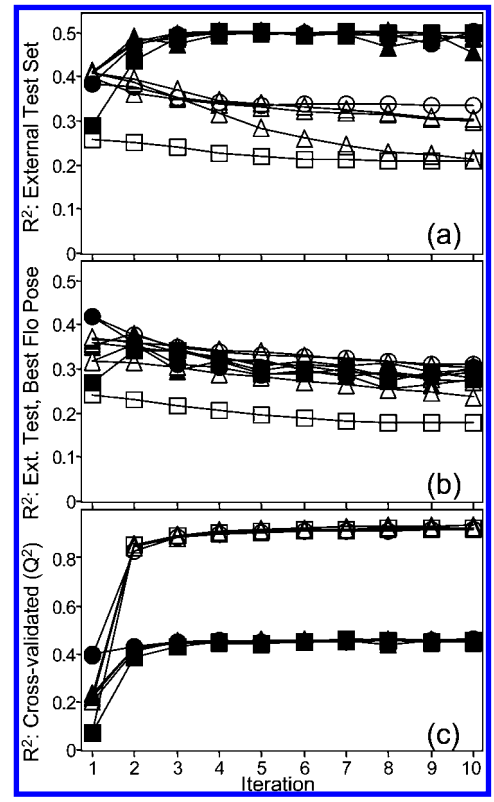

Figure 3. Evolution of PLS performance for CSF1R open crystal structure poses with Art-X descriptors. PLS models were initiated with the highest Flo+ scoring pose per compound (circles), the lowest Flo+ scoring poses (squares), and three sets of randomly selected poses (triangles). PLS models evolved by current-best pose selection (opaque symbols) and least-residual pose selection (open symbols). (a) Performance on the external test set of 25% withheld compounds. The correlation (Pearson, R^2) is between experimental pIC_{50} and the highest prediction from each test compound. (b) Correlation between pIC_{50} and the PLS prediction for the test set pose with the highest raw Flo+ score. (c) Correlation over training set fractions withheld in 5-fold cross-validation, i.e. Q^2.

in Flo+. Hand-built interaction features (see above) were extracted with Magnet. The 3D calculation is now finished. Each 3D pose having been reduced to a general-purpose docking score and a series of Magnet features. The remaining work is purely statistical modeling on these descriptors. The next step is iterative pose selection and PLS model building. The data set was randomly split 75/25 into training and test sets. The best-scoring (highest affinity) pose from Flo+ for each of the 1612 training set compounds was selected for the initial PLS model. The descriptors were the Magnet features, Flo+ score, contact atom counts, etc. (see the Methods section). The PLS model was built, and used to rescore all 76 719 training set poses. The new high-affinity pose for each compound from this rescoring was selected, and a second PLS model was built. This procedure was repeated 10 times to yield a final PLS model. The 25 891 poses from the 537 test set compounds were scored with the final PLS model. The activity for each compound was taken to be that of the highest scoring pose. The correlation of these predictions with experimental activity was $R^2 = 0.5$, substantially better than the success criterion of 0.32, taken from the best of the nearly 300 correlations from the extensive GSK study of conventional docking methods.[1]

Convergence and Stability of Iterative Pose Selection. Figure 3 takes a detailed look at iterative pose selection's

improvement to predictive R^2 (on the 25% held-out test set), at each iteration, for docking into CSF1R's open crystal structure, using the hand-built descriptors. In order to demonstrate the robustness of iterative selection of the lowest energy pose, model building was initiated using the best Flo+ scoring pose per ligand (the default method), the poorest Flo+ scoring pose, and three sets of randomly selected poses. After the first iteration, subsequent AutoShim models train on the maximum-predicted-activity (MPA) pose as usual. For comparison, results are also presented using the ill-advised least-prediction-residual (LPR) pose selection, in which the pose with a predicted activity closest to the measured experimental value is chosen to train the PLS model in each iteration.

The solid symbols in Figure 3a show that the MPA pose selection produces PLS models that converge in performance on the 25% withheld test set after just 2−4 iterations (Figure 3a). While, by default, poses with the best Flo+ score are selected to initiate PLS model building, apparently any pose selection yields equivalent performance. The best-Flo+ and random pose sets do yield better initial models than the worst Flo+ scoring pose, with initial R^2 on the 25% withheld test set of 0.38 (best-Flo+) and ~0.41 (random), while worst-Flo+ correlates at 0.29, but all have converged at about 0.50 after 4 iterations. The average predictive R^2 over iterations 6−10 for all five MPA pose selection trajectories is 0.49. The solid symbols in Figure 3c show that 5-fold cross-validated Q^2 on the training set converges similarly to R^2 on the 25% held out test set. Figure 3b shows that predictions based on the best Flo+ scoring pose degrade as the model and pose selection evolve to reproduce the affinity data better.

As mentioned in the introduction, a tempting but misguided approach to the multiple-pose problem would be, after each iteration, to choose the pose that most closely matches the experimental activity, rather than the best scoring pose by the current model., i.e., if the best-scoring pose overestimates the activity, choose a lower-scoring pose. The open symbols in Figure 3 show that this LPR pose selection is indeed a bad idea. Note that some other criterion besides LPR must now be used for predictions where the activity is not known. Two choices are available: the best score using the new PLS scoring function, or the best Flo+ score. The open symbols in Figure 3a and b show that predictions for the 25% withheld test set using the LPR PLS model, based on either the highest-affinity pose or the highest Flo+ scoring pose, degrade during evolution for all five starting points. The final models greatly under-perform the models derived using MPA pose selection.

Figure 3c demonstrates how badly one could be misled using LPR pose selection and monitoring performance of these overfit models by the cross-validation statistic, Q^2. The open circles show that 5-fold cross-validation for models built on LPR pose selection climb to $Q^2 > 0.92$. Recall that the MPA selection Q^2 values approximately mimic the external test set R^2, indicating that the normal procedure does not overfit.

Hopefully, the convergence in R^2 among MPA-evolved PLS models might indicate convergence to a common model, whereas the apparent overfitting of the LPR PLS models might indicate divergence. To test this, the PLS derived input variable coefficients from all 10 iterations of all 10 iterative pose-selection/PLS runs described above were combined, and projected to two dimensions using principal components analysis (PCA).[27] The projection explained 68% of the variance in the PLS coefficients. Lines connect the trajectories of each model to follow the progress evolution. The solid symbols in Figure 4a show that the MPA pose selection trajectories for different starting pose choices initiate from disparate regions in the coefficients-space, yet they do indeed converge to a common model. In sharp contrast, the LPR evolved PLS models do not converge, but wander in various directions, showing no stability in the final models. Figure 4b focuses in on the convergence of just the MPA pose selection models by repeating the PCA on the coefficients for just these models. The projection explains 55% of the variance. It shows in detail how the models all converge.

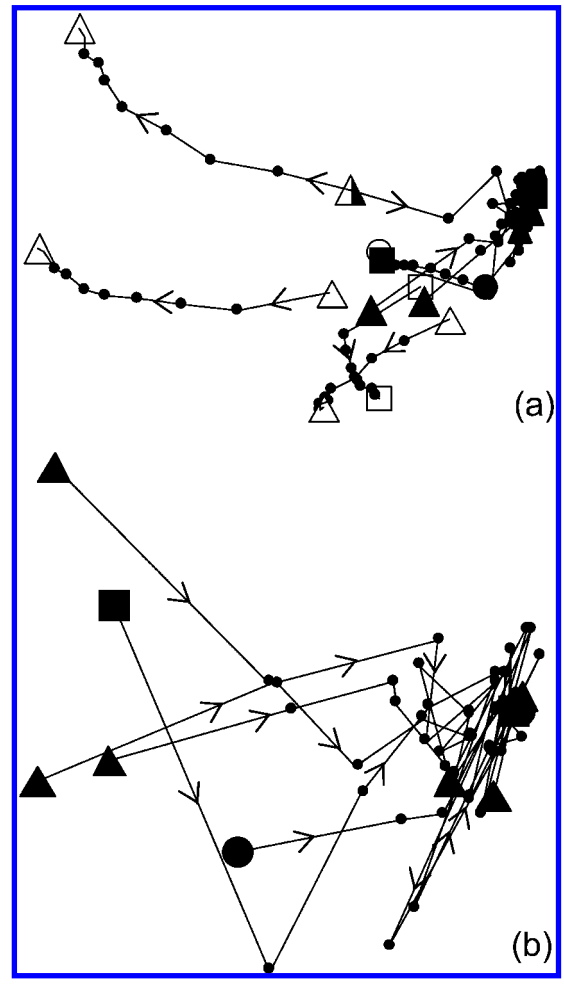

Figure 4. CSF1R PLS descriptor evolution. PLS models were initiated with the highest Flo+ scoring pose per compound (circles), the lowest Flo+ scoring poses (squares), and three sets of randomly selected poses (triangles), with current-best pose selection (opaque symbols) and least-residual pose selection (open symbols). The PLS derived descriptor weights were projected to two dimensions by PCA. Arrows indicate progression from iteration 1−10. In panel a, all 10 iterations from the 10 evolved PLS equations were subjected to PCA. Panel b magnifies just the five current-best-pose evolved PLS equations.

Adding Cross-Terms, Either Automated-RP or Expert-Built Art-X, To Improve Predictions. Table 1 compares the correlation between predicted and experimental pIC_{50} for AutoShim models derived from four different Magnet shim sets. Art refers to purely linear PLS on the hand-built ("art-based") Magnet H-bonding and pocket occupancy

Table 1. CSF1R AutoShim prediction accuracies (R^2) on a 25% Withheld Test Set, Using Four Different Shim Sets, with and without Cross-Term Generation by Recursive Partitioning (PLS-on-RP)[a]

model	Flo+	Art	Art-X	Auto	Art-X + Auto
			PLS[b]/PLS-on-RP[c]		
closed	0.04	0.25/0.43	0.26/0.35	0.39/0.49	0.42/0.46
open	0.28	0.43/0.49	0.50/0.52	0.46/0.60	0.50/0.60
open + closed	0.27	0.45/0.52	0.49/0.50	0.50/0.58	0.55/0.56

[a] PLS and PLS-on-RP predictive-R^2 results are averages over iterations 6–10 for numeric stability and to asses variances. Least significant differences associated with these variances, given by $(\sigma(2/5)^{1/2})t_4$, where $t_4 = 2.776$ (0.05 p-value), range from 0.002 to 0.026 R^2 units over the 24 AutoShim methods. [b] Number of descriptors in PLS models: Art, 20; Art-X, 63; Auto, 269; Art-X + Auto, 330. [c] Number of descriptors in final PLS-on-RP model (range over the three crystal structure sets): Art, 83–111; Art-X, 145–165; Auto, 169–197; Art-X + Auto, 181–208.

descriptors, Art-X refers to adding some selected quadratic cross-terms based on examining RP trees and noting the highly weighted descriptors in preliminary PLS equations, Auto refers to computer-generated pharmacophore-like Magnet shims, and Art-X + Auto combines the Art-based and automated shims. It further compares using features directly in PLS, or as data-driven, optimal, combination descriptors (PLS-on-RP), as well as comparing two different crystal structures, or an ensemble of both. The correlation between unshimmed Flo+ score and experimental pIC_{50} is also provided for comparison. See the Methods section for details.

In all cases, AutoShimming greatly improves predictions on the withheld test set relative to Flo+ score alone. Conventional docking into the closed crystal structure with just the all-purpose Flo+ scoring function gave no correlation with activity, but several AutoShim protocols gave acceptable activity predictions.

Applying PLS-on-RP significantly improves predictions relative to straight PLS on the Art shims, by 0.06 to 0.18 R^2 units. For the open and ensemble protein conformations, PLS-on-RP on the Art-X shims, which already include expert defined cross-terms, gives no further improvement. For the closed pose set, where traditional docking gave no correlation, handpicked quadratic terms did not help the correlation, but PLS-on-RP gave a huge improvement. Curiously, applying PLS-on-RP to the Art-X shims that already included handpicked cross-terms did improve over straight PLS, but less than applying it to the simple linear Art descriptors alone. The motivation for hybridizing RP with PLS was to capture the nonlinear dependencies among terms in the hand-curated Art-X descriptors, which appears to have been successfully realized. Not surprisingly, combining RP cross-terms with expert definitions fails to provide additional gains, suggesting that the two operations expose similar information.

Similar to straight Art descriptor based models lacking expert cross-terms, the simple automated pharmacophore shims likewise benefit from recursive partitioning cross-terms with predictive R^2 increasing by 0.08–0.14 units for the three CSF1R models. Overall, activity predictions from the fully automated method of PLS-on-RP applied to Auto descriptors are as good or better than any of the expert-influenced procedures.

PDK1. For PDK1, 1424 training and 489 test compounds with available IC_{50} data were docked into each of eight crystal structures. Correlations of straight Flo+ score with pIC_{50} for the eight individual docking models varied from $R^2 = 0.14$ to 0.25 (Table 2). Pooling poses from all eight crystal structures and selecting the highest Flo+ score for each compound irrespective of structure slightly improved the correlation to 0.27.

Table 2. PDK1 AutoShim with Eight Individual Crystal Structures and an Ensemble, Using Automated Pharmacophore-like Point Descriptors, with and without Cross-Term Generation by Recursive Partitioning (PLS-on-RP)[a]

docking model	Flo+	PLS	PLS-on-RP
individual (range of 8)	0.14–0.25	0.55–0.58	0.58–0.66
ensemble	0.27	0.60	0.64

[a] PLS and PLS-on-RP predictive-R^2 results are averages over iterations 6–10.

Based on the success of automated Magnet shims for CSF1R, art-based shims were skipped for PDK1. Ligand atom positions from the eight superimposed crystal structures were clustered to 66 feature positions. Four binary presence/absence pharmacophore descriptors, each with a 1.5 Å radius, were defined for each feature position: HBD, HBA, nonpolar, and aromatic. AutoShim scoring functions were built directly from these features as single-point shims (SPS) using straight PLS, and in optimal multipoint shims (MPS) using PLS-on-RP.

Again, the AutoShim scoring functions gave much better activity predictions than the all-purpose Flo+ scoring function. The variability in predictive accuracy over the eight single-structure PLS models (0.55–0.58) is tighter than the range in Flo+ correlations (0.14–0.25). We find no correspondence between Flo+ accuracy and PLS accuracy. In fact, the crystal structure with the best Flo+ score built the worst PLS model. Pooling all eight structures gives a slight accuracy improvement relative to the best single-crystal structure PLS model.

Table 2 shows that, as with CSF1R, adding RP generated MPSs consistently improved predictions over straight PLS on SPSs. The improvements for the eight single-structure models range from 0.02 to 0.10 R^2-units, with the best from the two classes improving by 0.08 R^2-units (Table 2). Pooling poses across the docking model ensemble yields a relatively good PLS-on-RP model, although slightly shy of best.

Application to Hit Finding. In an initial application, 100 000 compounds were docked into an AutoShim PLS-on-RP model using an ensemble of two PDK1 structures. The goal was to find new chemical scaffolds for this established target. The top scoring 1000 compounds were further evaluated by visual review of the dockings, by rescoring with a 2D Bayesian QSAR model, and with high priority for chemical novelty. Of the 23 compounds selected for assaying, 9 were active, with 5 submicromolar. The two most potent hits were the only two of the nine hits that were predicted to be active by both the 2D and 3D models. This suggests potential for consensus scoring among AutoShim and 2D methods.

Additional AutoShim Models: PLS-on-RP > PLS > Flo+. Table 3 lists performance of AutoShim models built for six other kinases using the automated descriptors. These models use a common set of shim definitions, derived by structurally aligning this ensemble of kinases, and clustering

Table 3. Additional AutoShim Models with Accuracies (R^2) on a 25% Withheld Test Set for PLS and PLS-on-RP

kinase	Flo+[a]	PLS[b]	PLS-on-RP[b]	data set size	$\sigma_{pIC_{50}}$
CHK1	0.12	0.53	0.57	2703	1.33
CSF1R (2x)[c]	0.22	0.39	0.44	2071	1.42
AurA	0.10	0.35	0.42	1108	1.11
GSK	0.02	0.36	0.40	4097	1.17
PI3K	0.13	0.37	0.36	1568	1.27
Tie2	0.00	0.30	0.24	382	0.68
PIM1	0.01	0.13	0.21	1145	0.82

[a] Flo+ R^2 calculated over highest score per compound with test and training sets combined. [b] PLS and PLS-on-RP predictive-R^2 results are averages over iterations 6–10. [c] Ensemble of two structures.

atom positions from a sampling of docked compounds to 75 points. The complete AutoShim descriptor set includes Flo+ score, atom count, and 300 binary shims consisting of 75 HBD, HBA, aromatic, and nonpolar features. Feature-radii are individually set as the distance to the nearest-neighboring feature center, which averages 2.3 Å and ranges from 1.1 to 4.3 Å. The dockings employ a sparser sampling of pose space (100 initial DockIt conformations prior to Magnet/Flo+ filtering), and are further reduced to a maximum of nine poses. Our accompanying manuscript[28] details this docking and pose-reduction. For reference, Table 3 also includes CSF1R. Its docking score accuracy ($R^2 = 0.22$) and number of passing poses, are slightly less with this reduced pose sampling, which may explain the low AutoShim R^2 values relative to Table 2.

Table 3 is ordered by performance using PLS-on-RP. Consistent with our other results, AutoShim substantially outperforms docking alone, and PLS-on-RP generally outperforms PLS. The two exceptions to PLS-on-RP's superiority are PI3K, where the difference is very small, and Tie2, which has a very small effective data set size due to a restrictive active site that filters out many compounds. Interestingly, performance (test set R^2) improves with a larger dynamic range in activity, as defined by the standard deviation in pIC_{50}. This trend is discussed further in our accompanying manuscript in the context of "Surrogate AutoShim".[28]

DISCUSSION

What is AutoShim? While AutoShim has been presented as a target-customized scoring function, it also has much in common with both 3D QSAR and 3D pharmacophore searching. In particular, like 3D QSAR and 3D pharmacophore searching, it requires training data. The quality of the AutoShim models depends on the quality of the training data. This is the biggest limitation of the method after the requirement for a crystal structure.

Overfitting is a potential problem in any empirically parametrized method, and it is worth summarizing the steps taken to minimize this risk. The first line of defense against overfitting is sufficient data. The compounds were diverse, not members of a QSAR series, and the data set sizes ranged from 382 to 4097, with an average of 1868. The number of parameters ranges from 268 to 302, so the data sets typically contain many times more observations than parameters. Furthermore, the robust PLS method is used for regression. PLS resists overfitting even when the number of parameters far exceeds the number of observations. Recursive partitioning is subject to overfitting, but it is only used to find cross-terms for the subsequent PLS, so overfitting would only select a suboptimal set of cross-terms. Pose selection happens outside the PLS and cross-validation steps, and is the most serious candidate for overfitting. This risk is minimized because the process converges in just 2 to 4 iterations, and our current standard protocol is now 5 iterations.

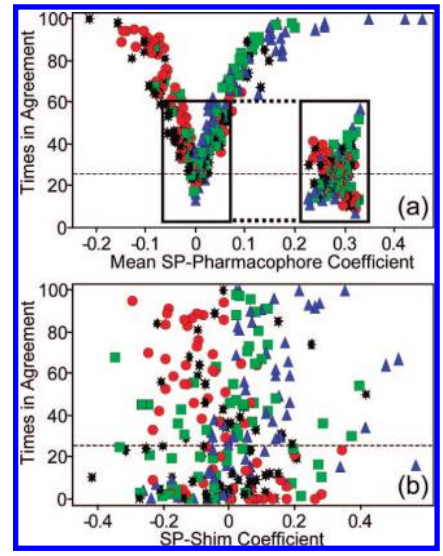

Figure 5. Coefficient sign agreement for PDK1. For the 100 50/50 data set splits used for SPP-AutoShim, the count of sign agreement between each SPP coefficient pair *and* the coefficient's mean value (averaged over all 200 PLS equations) is plotted against the mean coefficient value (a). The inset to panel a plots identical agreement-statistics for activity-scrambled data. Panel b plots the by-coefficient agreement-count between SPP pairs and the corresponding SPS coefficient. The dashed line at 25% is the random value for large numbers of splits. Values near zero indicate strong negative correlation. SPPs/SPSs are categorized as hydrogen bond donor (blue triangle), hydrogen bond acceptor (red circle), aromatic (green square), and nonpolar carbon (black star).

Numerous tests were employed to detect overfitting. Accuracy is estimated by testing the final model on a 25% held-out test set. The model coefficients were proven to converge to the same values irrespective of randomly selected starting poses as shown in Figures 3 and 4, and the large coefficients were insensitive to randomly subsetting the data set, as shown in Figure 5. These precautions and tests give us confidence that these models are capturing real interactions and not just fitting data.

AutoShim always begins with a general-purpose docking method. While this study used DockIt for docking and Flo+ as a docking score, any docking-program/scoring-function combination can be employed. Improved general purpose docking programs can only improve AutoShim. The innate accuracy of the docking model will determine whether the flavor of the AutoShim model is more "docking-like" or more "3D-QSAR-like." If docking alone accurately predicts activity, then the dock score should strongly inform the PLS equation, and AutoShim will act as a correction to the scoring function. If the general docking score does not predict activity, then the docking serves as a conformation filter and alignment tool, and AutoShim is more like 3D-QSAR.

Although not presented here, an array of scores using additional scoring functions can supplement the descriptor matrix, as well as scoring function components (e.g., Van

der Waals, electrostatics, solvation, strain). PLS optimally parametrizes their contributions, and RP can elucidate important nonlinear combinations among scores and components. The "consensus scoring" literature might suggest additional docking score combinations, such as average, minimum or maximum of rank-transformed scores among a privileged subset.[33,34]

High Predictive Power for Scaffold Hopping. The main advantage of AutoShim is its strong predictive ability across diverse compound sets. The predictions of $R^2 > 0.5$ are very competitive with 2D QSAR models, but they do not use any 2D connectivity information. Knowing that a new compound is topologically similar to a known active is a huge clue that it might also be active, especially for compounds with a similar scaffold. Conventional wisdom says that models that do not "know about" 2D topology should perform comparably on topologically unrelated compounds, and should therefore do better in "scaffold hopping" than 2D methods that give comparable correlations. There is little direct literature support for this generally held belief, and it is difficult to even imagine a systematic study that would confirm it. However, a recent paper that compiled "examples of successful scaffold-hops" indirectly supports it.[35] Only 2 of the 21 examples in their survey used strictly 2D methods.

Interpretation with Pure Single-Point-Pharmacophore (PLS-SPP) Models. Besides virtual screening, it would be useful if the shims were physically interpretable to assist further drug design. This is an advantage to using visualizable pharmacophore features as the shims. While the PLS-on-RP multipoint pharmacophore (MPS) models give the best activity predictions, interpreting these models is confounded for 3 reasons: (1) The models include the Flo+ score, so the pharmacophore shims are corrections to a general-purpose scoring function. (2) For every "foreground" rule in the recursive partitioning, such as "HBD near this point", there is a corresponding "background" rule i.e. "no HBD at this point" which does not describe an interaction (either positive or negative), but just collects the remaining compounds that lack this significant feature for further analysis. (3) The final prediction for any compound is a linear combination of many complex MPSs, which may include compensating and even contradictory interactions. To avoid these interpretation pitfalls, we built straight PLS models for PDK1 using only the 264 single-point pharmacophore (SPP) features, i.e. without including other continuous variables: the Flo+ score, nonhydrogen atom count, count of ligand atoms contacting protein, count of protein atoms contacting ligand, and ratio of receptor atoms contacting ligand to ligand atoms contacting receptor. In this case, the docking does not contribute directly to the predictive model, but only serves to generate the superimposed ligand conformations for what amounts to a 3D QSAR fitting. Since they only include point *pharmacophores* and no general-purpose docking score, they will be called PLS-SP*P* models to distinguish them from single-point shim (SPS) models, where the features are adjustments to a baseline docking score. The PLS-SPP model had a predictive correlation with experimental IC_{50} on the 25% held out test set of $R^2 = 0.56$, compared to $R^2 = 0.60$ for the PLS AutoShim model (PLS-SPS), and the 0.64 for the PLS-on-RP model that uses MPSs. While not as useful for virtual screening, the simple PLS-SPP model still has enough predictive power to expect interpretable coefficients.

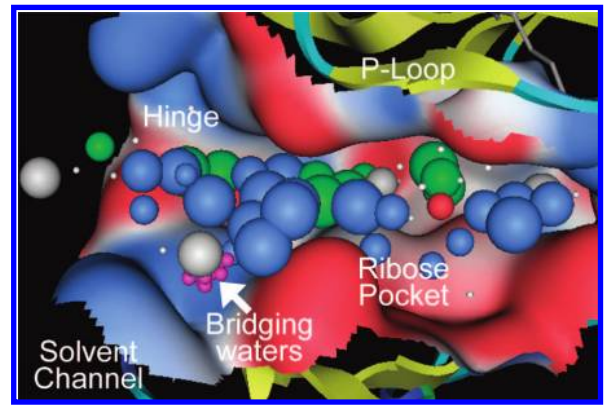

Figure 6. PDK1. Each probe position's most favored single-point pharmacophore (SPP) is color-coded as aromatic (green), nonpolar carbon (grey), hydrogen bond donor (blue), and hydrogen bond acceptor (red). Only SPPs with mean coefficient >+0.055 and coefficient sign agreement >57% are considered for display. If more than one SPP qualifies at a given position, the one with highest sign agreement is plotted. Larger balls are more significant. A small white ball indicates positions lacking a significant positive SPP. The positions of 15 waters, extracted from a structural alignment of 17 water-containing PDK1 crystal structures, are colored magenta. For reference, the electrostatic surface of a representative PDK1 crystal structure is shown over the ribbon backbone.

Stability of Coefficients in PLS-SPP Models. For the PLS coefficients to be interpretable, they must also be stable, i.e. the same pharmacophores must drive the affinity prediction equation irrespective of the specific compound training set. In order to assess this, the data set was split 50/50, 100 times, yielding 200 training sets. These 100 nonoverlapping pairs of PLS models were trained on just the 264 SPP descriptors. As a control, the same splitting procedure was repeated on scrambled activity data. Since all four pharmacophore types are included throughout the entire active site, most do not correlate with activity, and PLS assigns them small random coefficients in the regression equation. In a stable model, a small number of physically meaningful SPPs will dominate, reliably contributing the largest positive and negative coefficients, irrespective of the specific training set. In the scrambled data models, no pharmacophores should dominate beyond chance correlations.

Figure 5a and its shifted inset compare coefficient stability between the real and scrambled data by plotting on the *X*-axis the average coefficient for each SPP across the 200 training sets vs. on the *Y*-axis how often both coefficients for each SPP have the same sign as the mean in the 100 pairs of models from the 50/50 splits (which share no compounds in common). The rectangle in each plot contains all the points from the scrambled data, indicating the boundary of insignificant SPPs. The 7 most significant SPPs promoting affinity are all HBDs, followed by aromatics. Recall that only poses which form an H-bond with the hinge were included, so that alone does not explain this dominance. The most highly significant HBAs all have negative mean coefficients (decrease affinity).

Interpreting the Coefficients in PLS-SPP Models. Figure 6 shows, in the active site, the SPPs with the most significant positive coefficients (improve affinity). The sizes of the spheres indicate "significance" measured by sign agreement, and the colors show pharmacophore type. The model quite plausibly puts green aromatic features in the "aromatic pocket" and along the hinge where the adenosine

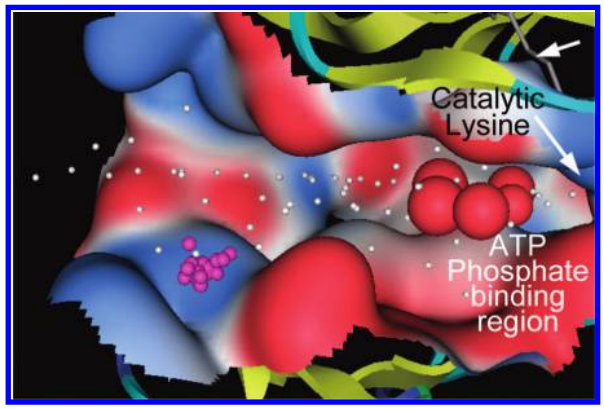

Figure 7. Hydrogen bond acceptor SPPs with a significantly positive mean coefficient value (>0.055) plotted as red balls in PDK1's active site. A white arrow labels the catalytic lysine's surface.

of ATP binds. Blue HBD features fill the solvent channel and the large negatively charged region covering the ribose and metal binding surfaces at the "bottom" of the pocket. Note the sparsity of significant H-bonding features near the hinge region. This is reasonable, because all retained poses have at least one H-bond to the hinge.

Surprisingly, several significant HBD features in the solvent channel are not within H-bonding distance of any HBA atom in the protein. However, a structurally conserved water bound to the backbone N of Glu166 in 15 of 16 PDK1 structures, which was removed before docking, is in position to accept these H-bonds. Thus the empirical parametrization appears to be inferring this interaction which was not in the actual docking model.

Interpreting which features are significant is tricky. Although individually their coefficients were each barely significant in Figure 5a, the 8 (of 66) most positive HBA features are all clustered along the phosphate binding region, and running back to the catalytic lysine (Figure 7). This must be more than coincidence. Presumably it is their combined weight across many poses that influences the overall fit.

Note that, unlike PLS-SPS AutoShim, which includes the Flo+ and ligand/protein contact scores, the PLS-SPP model is built without any knowledge of the protein. The consistency between the pharmacophore feature coefficients and the protein electrostatics, is entirely inferred by the PLS regression reproducing the ligand IC_{50} data. These sensible correspondences with the significant protein features lend credence to the supposition that PLS on the magnet features, combined with pose selection, captures physically meaningful interactions in ligand binding.

Comparing AutoShim and PLS-SPP Coefficients. Similar to Figure 5a, Figure 5b compares the AutoShim PLS-SPS model, which includes the Flo+ score and ligand/protein contact scores, with the PLS-SPP model which does not. It plots the PLS-SPS coefficient against the number of times that the coefficients in both pair-members of the 50/50 PLS-SPP splits agree in sign with the PLS-SPS coefficient. Note that random coefficients would lie at 25% (dashed line) for larger numbers of splits, and values near zero show strong negative correlations. Compared to the agreement within the PLS-SPP model shown in Figure 5a, it shows relatively little consistency between the SPP and SPS models. This confirms the hazards of directly interpreting shims, which are corrections to a standard scoring function, and endorses building a separate PLS-SPP model for pharmacophore interpretation and ligand design.

Examining PLS-on-RP Pharmacophores. While the PLS-on-RP multipoint models are confounded for several reasons, it is nevertheless instructive to inspect the top few multipoint shims (MPS) with the largest PLS coefficients. PDK1's PLS-on-RP eight crystal structure ensemble equation is composed of 207 terms, 78 of which are SPSs from decomposing the "seventh" tree.

RP MPS No. 1: Coefficient = 0.37. This MPS requires a HBD near the solvent channel, while excluding a HBD near the ribose binding region. An aromatic atom is required near the hinge, with no HBA at the gatekeeper. Also, the ligand must contain at least 25 nonhydrogen atoms.

The exclusion of the HBD in the ribose binding region is surprising, as it was beneficial in the PLS-SPP models, and would appear to benefit affinity given the many potential protein HBAs in that region. This rule is a background rule from an earlier foreground rule in the recursive partitioning tree that rewarded HBDs in this region. Such background rules are an expected complication in interpreting RP rules.

RP Rule No. 2: Coefficient = 0.36. The second most positive component in the PLS equation only applies to poses with a Flo+ score less than 5.65 and more than 25 nonhydrogen atoms. This does not mean a low Flo+ score improves activity; rather, it is another background rule. Flo+ > 5.65 appeared as a single-term rule from the initial and final split of an RP tree with coefficient of +0.22, as well as in three other multiterm rules, and as a "raw" unconditioned continuous variable. Flo+ with various thresholds also participates in 73 other multicomponent RP rules contained in the PLS equation.

The surprising aspect of rule no. 2 is a requirement for a nonpolar atom against the hinge. "Non-polar" is defined as a carbon covalently bound only to other carbons or hydrogen. In fact, several good inhibitors with in-house crystal structures place a nonpolar carbon within this volume, generally in an aromatic ring, and poses that do not form any H-bonds to the hinge had been filtered out by Magnet during the initial docking.

RP Rule No. 3: Coefficient = 0.35. Rule no. 3 is a single SPS for a HBD near the backbone carboxyl of Glu 90 in the glycine-rich P-loop.

Robust: Did Not Hand Curate Proteins or Ligands. AutoShim has proved to be a robust method, with the capacity not only to compensate for deficiencies in the general-purpose scoring function, but also other inadequacies such as the protein treatment: low-probability hydroxyl orientations, wrong assignment of nitrogen/oxygen in glutamine or asparagine, wrong histidine protonation, etc. QXP, which generated the Flo+ receptor models in this study, does some automatic optimization of hydrogens, but no further attempt was made to improve the X-ray structures. These nonetheless successful models suggest that shimming was able to compensate for a multitude of sins, including incomplete grooming of the protein structures.

Similarly, no attempt was made to identify structurally important waters. Instead, all waters were stripped before docking. As mentioned above, the second strongest AutoShim HBD feature in PDK1's solvent channel is not in the vicinity of any potential protein HBA atoms, but is adjacent to a strongly conserved water. This feature is also

consistent with the SAR. Without the AutoShim model, its impact might have been overlooked.

Iterative Pose Selection. Selecting MPA poses for iterative model building was applied earlier to align molecules for 3D-QSAR in the Compass program, which builds a neural network model of activity on distances to molecular surface probe points.[36] More recently, MPA selection been applied to 1D-QSAR, which requires a global (1D) molecular alignment.[37] They found, likewise, that optimizing on Q^2 led to unstable, overfit models. The success of this method signifies that additional studies should scrutinize the generality. Work in progress examines iterative pose selection using both linear regression and nonlinear methods with a discontinuous response surface such as recursive partitioning regression and random forest regression.

Throughput. The extra steps unique to AutoShim (Magnet feature extraction, RP, and iterative PLS) add very little computational time to the conventional docking on which it is based. Iterative pose selection does add computational time relative to PLS on a preselected set of poses, because all poses must be rescored over several iterations. Even so, PLS-on-RP using the Art + Auto descriptor set with poses pooled from two crystal structures takes only 5 h for 10 iterations. Figure 3 makes clear that only approximately four iterations are required, cutting the time in half. Further speedup in learning can be gained by clustering each compound's pose set. Over the entire training process, the slowest step is the conventional geometry optimization with Flo+, which take ~4 s per pose on a single processor, or ~100 CPU hours for the above ~100 000 poses. In practice, this was spread over a 190 CPU Opteron Beowulf cluster. Thus, training models is quick. Virtual screening takes longer, because so many more compounds must be docked. Screening an archive of 1 000 000 compounds takes about two weeks even on this large cluster. Even so, the ultimate slowest step is the assay development and experimental screening of the training and test set compounds.

CONCLUSION

High throughput docking is commonly employed for virtual screening, but general all-purpose scoring functions rarely correlate with measured binding affinity. Custom scoring functions trained on IC_{50} data have proven to be a practical remedy. Shims, similar to point-pharmacophores, can be added to the binding site of each protein target, using the program Magnet. Much like the shims in an NMR magnet, partial least-squares (PLS) regression adjusts the weights of the ligand interactions, correcting the scoring function to better reproduce IC_{50} data. Shimming dramatically improves the affinity predictions on 25% of compounds held out at random.

Docking produces many candidate poses for each ligand, each of which uses different interactions with the protein, and hence with the shims. Only the lowest-energy pose by the newly shimmed scoring function will make the appropriate interactions for affinity prediction, so this pose should be used to optimize the shims. However, this pose cannot be identified until the scoring function is trained. This "multiple pose problem" is solved by iteratively choosing poses and training the shim weights, always choosing the lowest energy pose by the current scoring function at each cycle, even if another pose would better fit the experimental activity data. This method reproducibly converges to a consistent solution in just 2−4 iterations, regardless of the starting pose or training compound set, so these robust models do not overtrain.

Three methods were used to generate shims: custom designing by hand to capture important interactions, automatically building unbiased shims using simple point pharmacophores, and automatically building biased shims using recursive partitioning to generate optimal multipoint pharmacophore shims. The best activity predictions come from scoring functions with over 100 complex multifeature shims generated by the recursive partitioning method, combined with a general purpose scoring function and whole molecule contact terms. However, these models are difficult to interpret when designing new compounds. Reasonable affinity predictions can still be obtained from sets of simple single-point pharmacophores, and these models clearly indicate the molecular interactions producing effective binding. The interaction requirements are very reproducible, irrespective of the compound sets used for training, lending confidence to the predictions and interpretations.

ABBREVIATIONS:

PLS, partial least-squares; RP, recursive partitioning; PCA, principal components analysis; LIE, linear interaction energy; HTI, high throughput intuition; HBA, hydrogen bond acceptor; HBD, hydrogen bond donor; SPS, single point shim; SPP, single point pharmacophore; MPS, multipoint shim; MPA, maximum predicted activity; LPR, least prediction residual.

REFERENCES AND NOTES

(1) Warren, G. L.; Andrews, C. W.; Capelli, A. M.; Clarke, B.; Lalonde, J.; et al. A Critical Assessment of Docking Programs and Scoring Functions. *J. Med. Chem.* **2006**, *49*, 5912–5931.
(2) Aqvist, J.; Medina, C.; Samuelsson, J. E. A new method for predicting binding affinity in computer-aided drug design. *Protein Eng.* **1994**, *7*, 385–391.
(3) Ortiz, A. R.; Pisabarro, M. T.; Gago, F.; Wade, R. C. Prediction of Drug Binding Affinities by Comparative Binding Energy Analysis. *J. Med. Chem.* **1995**, *38*, 2681–2691.
(4) Pan, D.; Tseng, Y.; Hopfinger, A. J. Quantitative structure-based design: formalism and application of receptor-dependent RD-4D-QSAR analysis to a set of glucose analogue inhibitors of glycogen phosphorylase. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1591–1607.
(5) Deng, Z.; Chuaqui, C.; Singh, J. Structural Interaction Fingerprint (SIFt): A Novel Method for Analyzing Three-Dimensional Protein-Ligand Binding Interactions. *J. Med. Chem.* **2004**, *47*, 337–344.
(6) Claussen, H.; Gastreich, M.; Apelt, V.; Greene, J.; Hindle, S. A.; et al. The FlexX Database Docking Environment - Rational Extraction of Receptor Based Pharmacophores. *Curr. Drug Discov. Technol.* **2004**, *1*, 49–60.
(7) Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D. Improved Protein-Ligand Docking Using GOLD. *Proteins: Struct., Funct., Bioinf.* **2003**, *52*, 609–623.
(8) Jansen, J. M.; Martin, E. J. Target-biased scoring approaches and expert systems in structure-based virtual screening. *Curr. Opin. Chem. Biol.* **2004**, *8*, 359–364.
(9) Magnet. http://www.metaphorics.com/products/magnet/index.html. (Accessed March 19, 2008).
(10) SYBYL. http://www.tripos.com. (Accessed March 19, 2008).
(11) Wold, H. Estimation of principal components and related models by iterative least squares. *Multivariate Analysis*; Academic Press: New York, 1966; pp 391–420.
(12) Chen, H.; Lyne, P. D.; Giordanetto, F.; Lovell, T.; Li, J. On Evaluating Molecular-Docking Methods for Pose Prediction and Enrichment Factors. *J. Chem. Inf. Model.* **2006**, *46*, 401–415.
(13) Perola, E.; Walters, W. P.; Charifson, P. Comments on the Article "On Evaluating Molecular-Docking Methods for Pose Prediction and

Enrichment Factors". *J. Chem. Inf. Model.* **2007**, *47*, 251–253.
(14) Antes, I.; Merkwirth, C.; Lengauer, T. POEM: Parameter Optimization using Ensemble Methods: application to target specific scoring functions. *J. Chem. Inf. Model.* **2005**, *45*, 1291–1302.
(15) Huang, S. Y.; Zou, X. Q. An iterative knowledge-based scoring function to predict protein-ligand interactions: I. Derivation of interaction potentials. *J. Comput. Chem.* **2006**, *27*, 1866–1875.
(16) Teramoto, R.; Fukunishi, H. Supervised Consensus Scoring for Docking and Virtual Screening. *J. Chem. Inf. Model.* **2007**, *47*, 526–534.
(17) Vieth, M.; Cummins, D. J. DoMCoSAR: a novel approach for establishing the docking mode that is consistent with the structure-activity relationship. Application to HIV-1 protease inhibitors and VEGF receptor tyrosine kinase inhibitors. *J. Med. Chem.* **2000**, *43*, 3020–3032.
(18) Huang, S. Y.; Zou, X. Q. An iterative knowledge-based scoring function to predict protein-ligand interactions: II. Validation of the scoring function. *J. Comput. Chem.* **2006**, *27*, 1876–1882.
(19) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893.
(20) DockIt. http://www.metaphorics.com/products/dockit.html.
(21) McMartin, C.; Bohacek, R. S. QXP: Powerful, rapid computer algorithms for structure-based drug design. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 333–344.
(22) Vieth, M.; Higgs, R. E.; Robertson, D. H.; Shapiro, M.; Gragg, E. A.; et al. Kinomics-structural biology and chemogenomics of kinase inhibitors and targets. *Biochim. Biophys. Acta, Proteins Proteomics* **2004**, *1697*, 243–257.
(23) ter Haar, E.; Walters, W. P.; Pazhanisamy, S.; Taslimi, P.; Pierce, A. C.; et al. Kinase chemogenomics: Targeting the human kinome for target validation and drug discovery. *Mini-Rev. Med. Chem.* **2004**, *4*, 235–253.
(24) Team, R. D. C. *R Development Core Team R: A language and environment for statistical computing*; R Foundation for Statistical Computing: Vienna, Austria, 2008. http://www.r-project.org/. (Accessed March 19, 2008).
(25) Wehrens, R.; Mevik, B. H. *pls: Partial Least Squares Regression (PLSR) and Principal Component Regression (PCR)*; R package version 1.2-0, 2006. http://www.cran.r-project.org/. (Accessed March 19, 2008).
(26) Dayal, B. S.; MacGregor, J. F. Improved PLS Algorithms. *J. Chemom.* **1997**, *11*, 73–85.
(27) Press, W. H.; Teukolsky, S. A.; Vetterling, W. T.; Flannery, B. P. *Numerical Recipes in C: The Art of Scientific Computing*, 2nd ed.; Cambridge University Press: Cambridge, 1992; p 1020.
(28) Martin, E. J.; Sullivan, D. C. Surrogate AutoShim: Predocking into a Universal Ensemble Kinase Receptor for Three Dimensional Activity Prediction, Very Quickly, without a Crystal Structure. *J. Chem. Inf. Model.* **2008**, *48*, 873−881.
(29) Goodford, P. J. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.* **1985**, *28*, 849–857.
(30) Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.* **1982**, *161*, 269–288.
(31) Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C. J. *Classification and Regression Trees*; Wadsworth: Monterey, CA, 1984.
(32) Therneau, T. M.; Atkinson, B.; Ripley, B. *rpart: Recursive Partitioning*; R package version 3.1-27, 2005. http://www.cran.r-project.org/. (Accessed March 19, 2008).
(33) Charifson, P. S.; Corkery, J. J.; Murcko, M. A.; Walters, W. P. Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *J. Med. Chem.* **1999**, *42*, 5100–5109.
(34) Wang, R.; Lu, Y.; Wang, S. Comparative Evaluation of 11 Scoring Functions for Molecular Docking. *J. Med. Chem.* **2003**, *46*, 2287–2303.
(35) Schneider, G.; Schneider, P.; Renner, S. Scaffold-Hopping: How Far Can You Jump. *QSAR Comb. Sci.* **2006**, *25*, 1162–1171.
(36) Jain, A. N.; Koile, K.; Chapman, D. Compass: Predicting Biological Activities from Molecular. Performance Comparisons on a Steroid Benchmark. *J. Med. Chem.* **1994**, *37*, 2315–2327.
(37) Diller, D. J.; Hobbs, D. W. Understanding hERG inhibition with QSAR models based on a one-dimensional molecular representation. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 379–393.

CI7004548