Exploiting Structure–Activity Relationships in Docking

David C. Sullivan* and Eric J. Martin

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

Received November 28, 2007

From the perspective of 2D chemical descriptors, error in docking activity predictions is separated into noise and systematic components. This error framework explains how fitting docking scores to a 2D-QSAR equation often improves accuracy as well as its logical limits. Intriguingly, in examined cases where multiple docking models (e.g., multiple crystal structures or multiple scoring functions) are available for an enzyme, the noise component of error dominates the difference between the more accurate and less accurate docking models. When this is true, the QSAR equation fit statistics can rank each docking-score set's accuracy in the absence of experimental activity data.

INTRODUCTION

Accurate computational methods to predict binding affinities for large numbers of ligands, given the structure of the target protein, (i.e., molecular docking), have eluded practitioners despite more than two decades[1] of development. Many studies have compared accuracy among docking programs,[2–6] analyzed sampling strategies,[7,8] and noted the physical shortcomings underlying molecular interaction models.[9]

This study takes a different tack in analyzing scoring errors by considering the statistical behavior of experimental activity values compared to error-prone estimates, including docking scores. One of the basic tenets guiding drug design is the "neighborhood principle", which states that chemicals of similar structure are more likely to share similar biological activity.[10,11] This implies a smoothness to the "activity surface" over an appropriate chemical structure space that permits identifying trends or quantitative structure–activity relationships (QSAR). While it has been noted that activity surfaces may exhibit roughness, with "small" chemical changes occasionally corresponding to large biological activity changes,[12] these "discontinuities" alternatively indicate the inadequacies in available chemical descriptors. An ideal activity surface can be defined as one smooth enough to fit with a QSAR equation. By definition, random errors defy structure–activity trends, adding discontinuity on top of what imperfect QSAR technology introduces. As explained below, these properties enable QSAR to evaluate and even filter out error in docking scores.

Although rarely advanced explicitly as a tool for error analysis, QSAR has danced with the subject over much of its history, generally focused on the related topic of how error enters predictions. Various factors complicate building accurate QSAR functions. First, the available set of chemical descriptors is vast.[13] Using all available descriptors is rarely a good choice. Soon after the advent of modern QSAR by Hansch and co-workers,[14] Topliss[15] recognized that, with limited data, some correlations between activity and descriptors are expected simply by chance without physical relevance. Including too many descriptors often results in an equation that reproduces random error and idiosyncrasies inherent in a data set but lacks predictive power. In practice, QSAR model complexity can be tamed by eliminating descriptors with small variance over the training set, eliminating strongly correlated descriptors, and constraining function parameters, for example by eliminating low-variance latent variables in partial-least-squares regression (PLS). Second, the sparse and/or limited sampling of the activity hypersurface limits the accuracy of a QSAR model in predicting activities of untested compounds. While an ambitious Big Pharma high throughput campaign may screen $\sim 10^6$ compounds, the potential number of small molecules conservatively exceeds 10^{50},[16] Third, the *measured* activity surface may be distorted by assay errors. Finally, the chosen functional form of the QSAR equation may not accommodate an accurate reproduction of the activity surface.

In response to QSAR's imperfections, practitioners are satisfied to model only the larger activity trends in a data set to avoid fitting biophysically irrelevant features of the training data, necessarily smoothing out some of the signal. Cross-validation[17,18] can optimize this balance. The hypothetical QSAR fitted to experimental activity in Figure 1 (top scheme) illustrates the error associated with this smoothing. Imperfect QSAR technology thus imposes a "neighborhood limit" to QSAR accuracy (see below).

Training on error-prone activity estimates such as docking scores can attenuate some errors but introduce new errors into the QSAR equation, as illustrated in Figure 1 (bottom scheme). First consider the left-hand side, where errors appear to be random with respect to the chemical axis. We term these errors as "non-neighborhood", meaning that they do not depend on proximity in chemical space. Fitting to a QSAR equation filters away non-neighborhood errors, potentially improving accuracy up to the neighborhood limit. But, notice how the activity estimator falsely reports all compounds projecting to the right-hand of the chemical axis as inactive. This illustrates a "neighborhood error", meaning

* Corresponding author phone: (510) 923-3306; fax: (510) 923-2010; e-mail: davidxyz1972@yahoo.com.

10.1021/ci700439z CCC: $40.75
Published on Web 04/10/2008

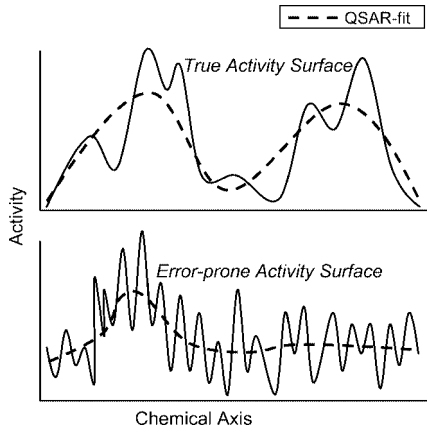

Figure 1. Hypothetical QSAR training on activity and error prone activity estimate. A hypothetical scenario illustrates QSAR smoothing. The top panel plots the actual (experimental) activity as a function of the chemical structure axis (solid line), with the QSAR fit to a sample of compounds plotted as a dashed line. The imperfect QSAR fit reflects the neighborhood limit of an appropriately constrained QSAR equation. The bottom panel plots an error prone activity estimate (solid line) along with a QSAR equation fit to a sample of estimates. The estimates contain non-neighborhood errors throughout the chemical space, while a neighborhood error afflicts the right-hand side active chemotype.

that activity is systematically misestimated over the entire chemical neighborhood. Neighborhood errors cannot be removed by QSAR-fitting.

As practical examples, QSAR-fitting has been applied to error-prone high throughput activity measurements[19] as well as molecular docking activity predictions[20] to improve recovery of experimental true actives. Krumrine et al.[21] employ a similar logic in reranking docking hits as informed by the docking-hit distribution in their immediate chemical neighborhood; docking hits from clusters with many other hits are presumed more likely to be true positives than docking hits from poorly scoring clusters. These strategies can only succeed in improving docking results in the presence of non-neighborhood error. These successes might be surprising, given the potential for significant systematic errors in docking. For example, neglecting ligand conformational entropy may lead to systematic overestimates of flexible ligand activity, while neglecting desolvation may yield activity overestimates for charged ligands. Neglecting protein flexibility may bias chemotypes similar to the crystallized example. Additional approximations in scoring functions will add additional reproducible errors. Chemical descriptors will fit some of these errors. Other errors may be reproducible yet not map onto the chosen chemical descriptors in a smooth fashion and hence classify as non-neighborhood.

Given that error breakdown into neighborhood and non-neighborhood limits the improvement QSAR-cleaning can impart on docking, this study places considerable attention to characterizing error structure. Prior to examining docking scores, QSAR is used to investigate the relationship between the "neighborhood limit" and "noise reduction" in activity data sets containing synthetic random errors. This limiting scenario provides baseline references that support insight into the more complex error structure in docking and help rationalize docking accuracy improvements achieved by QSAR-cleaning.

Finally, training predictive models (i.e., 2D-QSAR) on predictions (i.e., docking) greatly expands the possibilities for virtual screening workflows relative to general QSAR applications. Specifically, assay data availability no longer limits chemical diversity of the QSAR training set, unleashing training set design as an important task for the computational chemist. We explore a handful of methods in this arena based on target-specific docking scoring functions and offer some guiding principles.

THEORY AND METHODS

Data Sets. The experimental data sets for kinases are proprietary multiconcentration inhibition measurements (IC_{50}). Data modeling uses the negative-logarithm of IC_{50} (pIC_{50}). Activity data sets for 9 protein kinases used for noise-added analysis range in size from 709 to 8005 compounds. An additional three kinase compound sets with pIC_{50} measurements against kinases with multiple available crystal structures were used in docking studies. These compound set sizes are 2165 (CSF1R), 1913 (PDK1), and 1576 (PIM1). However, only compounds that successfully dock (see below) to all crystal structures in their enzyme set were analyzed, reducing the effective compound set sizes to 1988, 1498, and 1569, respectively. The standard deviations in pIC_{50}, a measure of dynamic range, are 1.4 (CSF1R), 1.3 (PDK1), and 0.7 (PIM1). These data sets contain a diversity of chemical structure with many unique "Bemis and Murcko" molecular frameworks,[22] in which side chains have been stripped off leaving only ring systems and linkers. Disregarding atom types, hybridization, and bond order, the three docking data sets contain 812 (CSF1R), 522 (PDK1), and 556 (PIM1) 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. Finally, 70 kinase data sets, including the above 12, were used for training target-tailored scoring functions for docking, that provide additional activity estimates.

In addition to kinase activity data, an HIV-1 protease pIC_{50} data set of 1015 unique compounds was obtained from BindingDB.[23] Multiple activity values reported for a single compound were averaged. This compound set contains 254 unique molecular frameworks.

Noisy pIC_{50} Data. Noisy activity data for QSAR fitting were generated using three error models: uniform error, normal error, and *y*-scrambling. Uniform error was applied by adding a uniformly random deviate on $[-d, d]$, where d assumes the values 0.25, 0.5, 1, 2, 3, 4, 6 and finally pIC_{50} is replaced by a random number on [0,1]. Analogously, zero-mean normal error deviates draw from normal (Gaussian) distributions with standard deviations defined by the same set of d values. For *y*-scrambling, a percentage of the compounds (5, 10, 20, 30, 40, 50, 60, and 100%) had their pIC_{50} values randomly reassigned among the subset.

2D-QSAR. This study primarily uses QSAR for smoothing docking scores in order to filter noise. Given that the objective is not to predict docking scores, per se, typical QSAR model selection techniques including cross-validation do not apply. Instead, parameters were fixed and biased by experience to lie in the vicinity of optimal values from previous studies. Specifically, we mainly employ linear partial least-squares regression (PLS)[24] with 10 latent

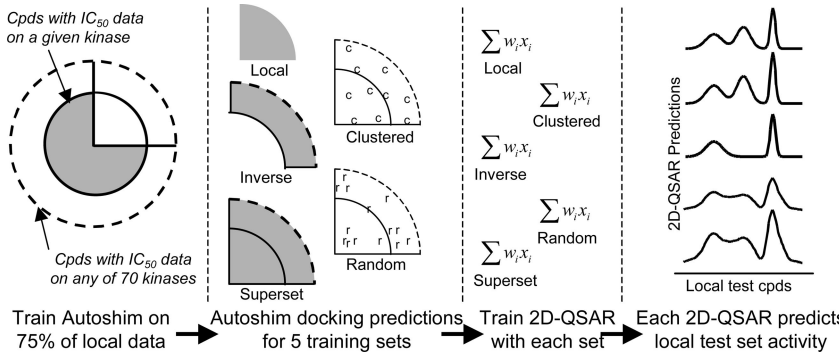

Figure 2. Training set design for AutoShim QSAR-cleaning.

variables; however, other latent variable choices were employed as well as ordinary least-squares (OLS) regression. We exclusively employed Scitegic's[25] 2D extended-atom-connectivity fingerprints with functional atom types (FCFP) as chemical descriptors, most often with a six-bond diameter (FCFP6), although where specified we also use the two-bond diameter fingerprint which carries less information. Poorly sampled 2D fingerprint features were excluded by means of an occurrence-threshold, which is generally equal to the square-root of the compound set size; however, we also use a ten-count threshold where noted. Regressing against all populated FCFP6 features, which often number more than 10,000 across our data sets, was computationally infeasible. The occurrence-thresholds generally reduce the number of features to below 1000. All 2D-QSAR was performed using Scitegic's Pipeline Pilot.[25]

Conventional (Untailored) Docking to Kinases. Docking into kinases used internally solved crystal structures for CSF1R and PDK1 and PDB[26] structures for PIM1. PDB IDs and chains used for PIM1 docking are as follows: 1XQZa, 1XR1a, 1XWSa, 1YHSa, 1YI3a, 1YI4a, 1YWVa, 1YXSa, 1YXTa, 1YXUa, 1YXUb, 1YXUc, 1YXUd, 1YXVa, 1YXXa, 2BIKb, 2BZJa, 2BZKb, and 2C3Ib. Docking proceeded by generating an initial set of 250–400 poses (fixed for each kinase) in the ATP binding site using DockIt.[27] Magnet[28] was used to eliminate poses without at least one hydrogen bond to the hinge connecting the N- and C-terminal kinase domains, biasing sampling toward kinase relevant poses, as known kinase inhibitors generally adopt one of the three hydrogen bonds which stabilize the adenine ring of bound ATP.[29,30] Passing poses were energy minimized and scored with the Flo+ component of QXP+.[31] The Flo+ score combines molecular mechanics components and empirical solvation terms and is parametrized to estimate relative pIC_{50} values. For each compound, the highest Flo+ score among all passing poses defines its "Untailored" docking-predicted pIC_{50}.

Conventional Docking to HIV-1 Protease. HIV-1 protease docking models derived from 24 public crystal structures deposited with the protein data bank.[26] The PDB IDs are as follows: 1D4H, 1EC1, 1FQX, 1GNO, 1HIH, 1HPV, 1HVI, 1HVJ, 1HVK, 1HVL, 1M0B, 1NPA, 1NPW, 1OHR, 1U8G, 1W5W, 1W5X, 1WBK, 1XL2, 1XL5, 1ZSF, 2AQU, 2FDE, and 7UPJ. They were selected to cover the three most populated crystal forms, associated with space groups P61, P21212, and P212121, drawing (arbitrarily) eight structures from each. Structural waters identified by DockIt's site preparation utility with default parameters were retained for docking and minimization. For all but four structures, only one water was retained, located between the flaps. In two of the four exceptions (1W5W and 1W5X), one additional water remained hydrogen bound between Arg8, Gly27, and Asp29. The other two exceptions (1XL2 and 7UPJ) had all waters stripped. DockIt and QXP+ were employed for docking, minimizing, and scoring. DockIt initially generated a pose set of 125 poses, of which the top 50 were carried forward to Flo+ minimization.

"Surrogate AutoShim": Target-Tailored Scoring Functions for Kinases. In addition to examining error in conventional docking scores, we also investigate predictions generated by empirically parametrized, target-tailored surrogate docking scoring functions. Surrogate AutoShim uses PLS regression to add "shims", i.e. pharmacophoric features, to the Flo+ scoring function from docked poses across a "universal kinase" ensemble of 16 diverse kinase crystal structures.[32] Briefly, recursive partitioning of the individual features generates multipoint pharmacophore "shims". These multipoint shims are treated as binary descriptors that are added to the original docking score and regressed against a random 75% of the experimental pIC_{50} values by PLS to generate the target-tailored scoring function.

A fitting procedure was already used to train the AutoShim docking model for the initial 75% of the data for each kinase. Hence, the 25% withheld set of compounds with pIC_{50} data on each kinase, which defines the "local" Autoshim test set, are held out to use for testing subsequent 2D-QSARs (and sometimes for training further 2D-QSAR models). Training the 2D-QSAR on AutoShim/docking predictions also allows training 2D-QSAR models on compounds with no experimental assay data and introduces training set design as an important component to restating 3D-models with 2D-QSAR. We utilize five 2D-QSAR training sets for each of the 70 kinases as illustrated in Figure 2. The 5 sets were assembled as follows:

(1) The "local" test set, a random 25% of experimental IC_{50}s for each given kinase (see above).

(2) A "superset" of all 18,766 unique local test set compounds from all 70 assayed kinases. The "superset" 2D-QSAR models permit examining whether adding extra compounds lacking activity data washes out the signal or improves stability by increasing chemical diversity. Unlike the local test sets, this compound set is the same for all 70 kinases.

(3) The "superset" was clustered using Pipeline Pilot with FCFP6 descriptors to 3787 compounds yielding the "clustered" set. Again, the same for all 70 kinases.

(4) An additional set of 3787 compounds was randomly selected from the superset. Whereas the "random" set should

have a substructural distribution similar to the superset, clustering enforces a more uniform distribution in chemical space. Clustering also reduces the fraction of actives, which tend to cluster together in our data sets. Again, the same for all 70 kinases.

(5) Finally, subtracting the "local" sets from the "superset" defines the "inverse" sets – training sets specific to each kinase. Performance on the kinase's 25% withheld experimental assay results with an "inverse"-trained 2D-QSAR measures two levels of predictive ability in that the ultimate test compounds neither inform the original AutoShim docking model nor the 2D-QSAR equation.

For consistency, testing of the final 2D-QSAR is always performed on the same 25% withheld "local" set for each kinase, irrespective of the 2D-QSAR training set used.

RESULTS

The error structure in docking-predicted activity is likely to be complex and originate from myriad sources. Hence, we begin with investigations using controlled error, manufactured by explicitly adding random noise to experimental activity data. We quantify the drop in both activity prediction power and the 2D-QSAR equation's fit to the training data with increasing noise. These "toy model" results provide baselines for investigations into the error structure of docking scores.We demonstrate the utility of 2D-QSAR in attenuating error in both untailored and target-tailored docking activity predictions. We also show that "stronger SAR" throughout docking scores, as measured by the 2D-QSAR fit to a compound set's scores, appears to be associated with higher accuracy, as measured by the correlation between docking scores and experimental activity. Neighborhood and non-neighborhood error is invoked to understand this result.

QSAR Models of Noisy Activity Data. Multiple noisy data sets for each of the nine kinases were constructed using three error models (uniform error, normal error, and *y*-scrambling) as outlined in the Theory and Methods section. The amount of true signal remaining in these noisy data sets ("data set accuracy") is quantified by the square of Pearson's correlation, R^2, with the original noise-free pIC_{50} data set. R^2 measures the fraction of overlapping variance between two distributions. Error can therefore be defined by $1 - R^2$. For each of the 9 kinases, a total of 24 noisy data sets were constructed. Each noisy data set, plus the original noise-free data set, independently trains a QSAR model by PLS with 10 latent variables using those FCFP6 features occurring at least $\sqrt{N}$ times, where N is the number of compounds in the data set.

Following QSAR modeling, two quantities are of interest. First is the "QSAR-fit", defined by the correlation (R^2) between the (noise-added) training data and their QSAR estimates. "QSAR-fit" should not be confused with our second quantity of interest, "QSAR accuracy". QSAR accuracy is defined by the correlation (R^2) between the original noise-free pIC_{50}s and the same QSAR predictions. In the special case where noise-free pIC_{50} data train the QSAR equation, QSAR-fit equals QSAR accuracy.

Figure 3 plots data set accuracy, our inverse measure of noise, against QSAR-fit. When plotted in this fashion, QSAR-fit degrades by a simple linear relation with noise independent of the error model (uniform, normal or

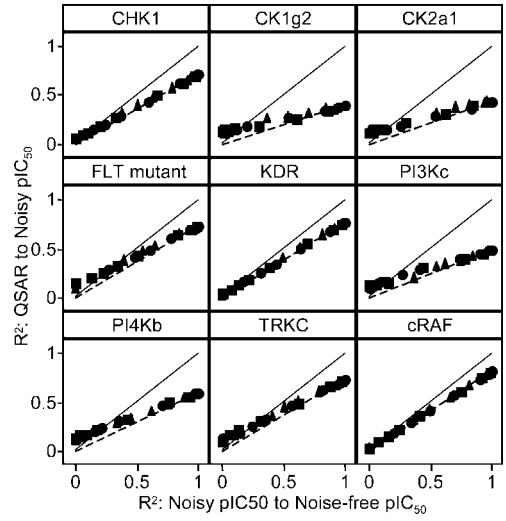

Figure 3. Synthetic noise: QSAR-fit vs added noise (data set accuracy). QSAR equations were built for nine series of data sets. Each series derives from pIC_{50} values for the kinase denoted. Each data set had error added prior to building the QSAR equation by adding a uniform random deviate (squares) or a normal deviate (circles) or by scrambling a percentage of the pIC_{50} values (triangles). The correlations (R^2) between the training data and QSAR equation values are plotted as a function of the correlations (R^2) between training data and noise-free pIC_{50}. An additional data point with no error (data set accuracy = 1) has a square symbol. The dashed line plots the idealized QSAR-fit response to error with a very large N (see the Appendix). For all figures, the sign of R is applied to R^2 to indicate cases of negative correlation.

y-scrambled) all the way from no added noise ($x = 1$) to pure noise ($x = 0$). For an infinite population, the linear relation would be perfect, from the R^2 of the noise-free QSAR to the origin (see the Appendix). A *dashed* line plots this ideal relationship in Figure 3. If the QSAR perfectly fit the noise-free pIC_{50}, it would follow $x = y$ (solid lines in Figure 3). For these real data, least-squares lines drawn through the nine sets of QSAR-fit values in Figure 3 (regression lines not plotted) all fit with $R^2 >= 0.978$. For the three data sets with greater than 2000 compounds, R^2 exceeds 0.999. The *y*-intercepts (correlation with pure noise) range from 0.03 to 0.14, with larger compound sets intercepting closer to zero. Noise-free pIC_{50} data sets yield stronger fits where there is greater dynamic range, with QSAR-fit and the variance across pIC_{50} data rank-correlating at 0.9. In sum, these trends should not surprise practitioners of the art. This simple linear plot, independent of the random error model, motivates using QSAR as a tool for analyzing error in activity measurements or predictions (below).

Figure 4 reanalyzes the QSAR equations from Figure 3 by plotting QSAR accuracy against data set accuracy. The ideal and limiting QSAR-fit lines from Figure 3 are carried to Figure 4. In contrast with QSAR-fit, the QSAR equation's correlation to the original noise-free pIC_{50} degrades relatively slowly with increasing noise (i.e., moving from complete to nil correlation on the *x*-axis). This means the QSAR's ability to predict the actual experimental activity of new compounds can be much higher than the fit statistics of the QSAR method would indicate. In six of the nine cases, QSAR prediction accuracy actually exceeds data set accuracy over part of the noise regime (points above the $x = y$ diagonal in Figure 4). The noisier the data, the stronger this effect. In these cases, QSAR estimates from fitting data with substantial random

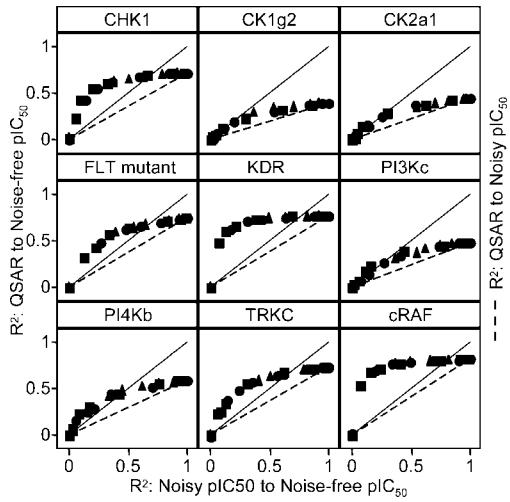

Figure 4. Synthetic noise: QSAR accuracy vs added noise (data set accuracy). pIC_{50} predictions calculated by the QSAR equations trained on noise-added data (see Figure 3) were correlated with experimental pIC_{50}. These correlations are plotted as a function of the correlation between noisy-pIC_{50} values and experimental pIC_{50}. The thin solid line plots the $x = y$ diagonal, for reference. The dashed line plots the idealized QSAR-fit response to error with many compounds, taken from Figure 3 (right-hand axis, identical scale) (see the Appendix).

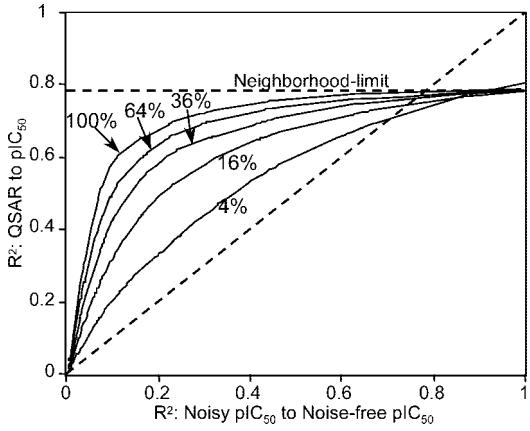

Figure 5. QSAR accuracy dependence on data set size. The KDR activity data set's QSAR accuracy dependence on data set accuracy (see Figure 4) was recalculated using random subsets of 4, 16, 36, and 64% of the entire data set. For each percentage, the QSAR accuracies were smoothed and are plotted as a continuous curve, for presentation purposes.

noise correlate better with the actual (noise-free) pIC_{50} than the noise-added data themselves. Presumably, this is due to the QSAR enforcing neighborhood behavior. This suggests the use of QSAR "smoothing" to clean very noisy activity data, a facility recognized in previous studies[19,20] and further demonstrated below. This amount of noise is generally greater than most experiments but is well within the range of docking estimates (see below).

The drop in QSAR accuracy upon adding noise likewise depends largely on the number of compounds being QSAR smoothed. Figure 5 replots QSAR-accuracy for the largest data set from Figures 3 and 4 as well as for subsets chosen randomly. QSAR accuracy for larger data sets falls more slowly upon noise addition than for smaller data sets.

The linear drop in "fittable" signal with noise forms the basis for a model to assign sources of accuracy. The vertical line in Figure 6 dissects a data set with accuracy of 0.34,

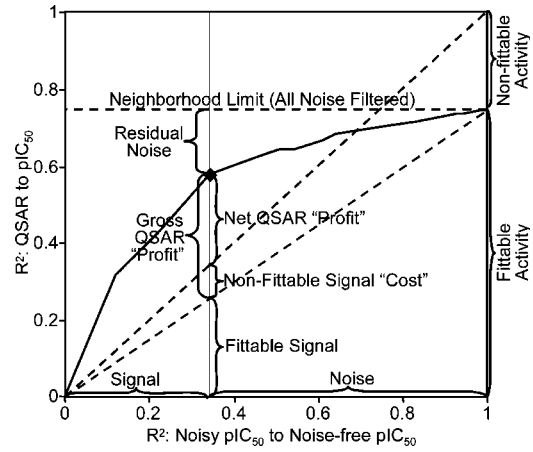

Figure 6. Error accounting in QSAR trained on noise-added data sets. A noise-added FLT3 mutant activity data set (data set accuracy = 0.34) has been QSAR-smoothed, yielding a QSAR accuracy of 0.58. The bow-line plots QSAR accuracy for the full range of data set accuracies, approaching the neighborhood limit of 0.75. The neighborhood limit interpolates to 0.26 ("fittable signal") at data set accuracy of 0.34. The QSAR accuracy improvement above fittable signal (0.58 – 0.26 = 0.32 R^2 units) defines the "gross QSAR profit". Nonfittable signal (0.34 – 0.26 = 0.08 R^2 units) is a "cost" that reduces the "net" profit to 0.24 R^2 units, i.e. prediction accuracy minus data set accuracy.

i.e. R^2 of noisy data to pIC_{50} is 0.34. The "Net QSAR Profit" is the improvement in estimating the activities of noisy data by QSAR smoothing. If all the non-neighborhood error were removed by QSAR, as in a very large data set, then QSAR accuracy would approach the "neighborhood limit" associated with the QSAR equation, 0.75 in this case. Applying QSAR to the noisy data set introduces a "cost", because one-fourth of the activity signal is unfittable in the noise-free data, as well as a "profit" from reducing noise through QSAR smoothing. The "gross-profit" adds to the fittable portion of activity signal. The difference between the noise-reduction "profit" and the neighborhood-limit "cost" is added to the data set's 0.34 accuracy. Of the 0.49 R^2-units potential gain in accuracy from QSAR smoothing (difference between neighborhood limit of 0.75 and the fittable signal component, 0.26), 0.33 correlation units are actually realized, or 66%. Figure 5 suggests that additional sampling over the chemical space would increase this percentage. While this high-noise data set net-profits from QSAR cleaning, where the "bow" curve of QSAR accuracy crosses the $x = y$ diagonal (0.70), profit exactly equals cost. For this target, data sets containing less than 30% noise net-degrade in accuracy with QSAR smoothing.

Figure 6 might mislead one to apply very complex QSAR equations with the best fit to pIC_{50} for QSAR smoothing, as this minimizes the cost associated with unfittable signal. However, the other term of the net-profit equation, noise reduction, will suffer as QSAR complexity increases. In order to examine the optimal balance between noise-smoothing and signal retention as QSAR complexity varies, QSAR accuracy was calculated for noisy data sets smoothed by PLS employing different numbers of latent variables (Figure 7). As expected, the more complex QSAR equations, containing more latent variables, better fit any given noisy data set (Figure 7, part a). The linear relationship seen in Figure 3 persists when applying a different number of latent variables, with correlation (R^2) between QSAR-fit and data set accuracy

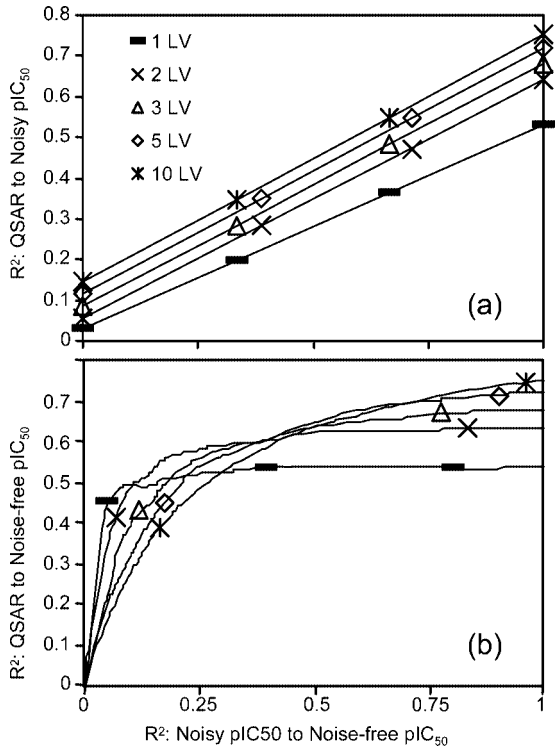

Figure 7. Synthetic noise: PLS latent variable dependence. FLT3 mutant activity and noise-added-activity were QSAR modeled by PLS using 1, 2, 3, and 5 latent variables, in addition to 10 latent variables as in Figures 3 and 4. For each latent variable value, panel (a) plots QSAR-fit vs data set accuracy, whereas (b) plots QSAR-accuracy's trend with data set accuracy. For presentation purposes, best-fit lines (a) and trend lines (b) plot the relationship for each latent variable and are based on more noise-added data sets than presented in Figures 3 and 4. QSAR-fit and accuracy values cited in the text correspond to actual data sets.

exceeding 0.992 for the five latent variable choices examined (Figure 7, part a). However, the prediction accuracies show a more complex relationship with the number of latent variables (Figure 7, part b). In the limit of zero added error, including more latent variables invariably improves the fit to the activity signal. However, with added error, too much complexity in the QSAR equation diminishes activity surface "smoothing". The additional signal captured by the more complex QSAR model does not fully compensate the additional random error that is fit as well. The optimal tradeoff between error smoothing and activity-signal preservation shifts to more parsimonious QSAR equations in the high-random-error regime, with the one-latent-variable model performing best at predicting pIC_{50} where the noise component exceeds 85% (data set accuracy <0.15, Figure 7, part b). While Figure 7 only plots results for one kinase data set, calculations performed for the remaining 8 data sets in Figures 3 and 4 all show similar trends.

Neighborhood vs Non-Neighborhood Errors: Docking. Training 2D-QSAR on 3-D docking is potentially useful for analysis of docking errors, QSAR-cleaning of docking results to improve prediction accuracy and also for training computationally less demanding surrogates for expensive docking calculations. Similar to the manufactured noisy data examined in Figures 3 and 4, molecular docking also offers error-prone estimates of activity. In a recent comparison of 10 high throughput docking programs applied to 8 targets using a total of 37 scoring functions, the maximum correlation with

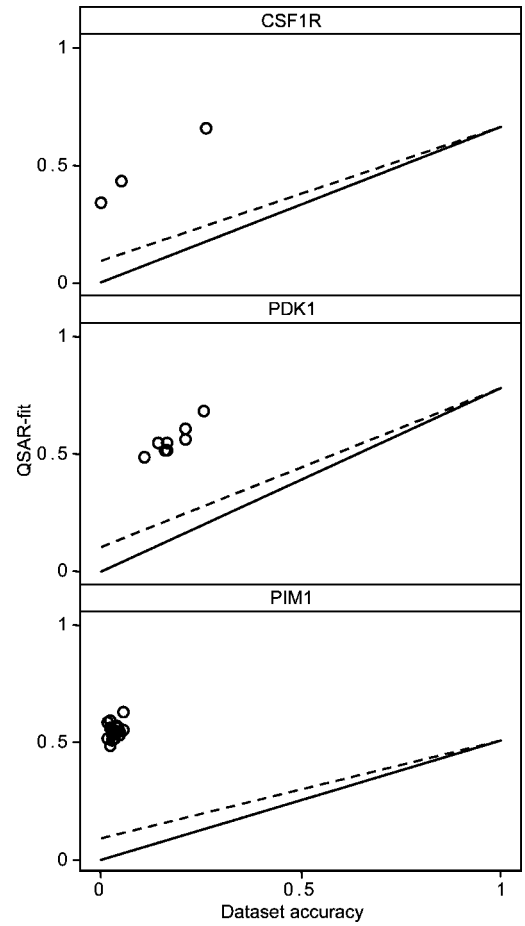

Figure 8. Docking: QSAR-fit vs data set accuracy. QSAR-fits to docking scores using many crystal structures for each kinase are plotted as a function of data set accuracy (i.e., docking score correlation to pIC_{50}). To compare docking score error behavior to pure noise, the solid line plots the idealized QSAR-fit response to added noise, while the dashed line plots the actual response.

pIC_{50} was 0.32.[2] The remaining ≥68% of the variance in docking scoring contains some ratio of neighborhood to non-neighborhood error in a given chemical descriptor space. QSAR-fits offer one operational means for their distinction. Figure 8 analyzes docking-based activity predictions from multiple crystal structures for three kinases: CSF1R, PDK1, and PIM1. I.e., for each crystal structure, the activity estimates from docking into that structure are correlated with measured pIC_{50} to get the data set accuracy. The QSAR fit *to the docking data* is then plotted against that data set accuracy to generate a single point on the plot. QSAR-fits as a function of data set accuracy for the previously described noise-added pIC_{50} data provide a reference for QSAR equations trained on docking scores. The QSAR-fits for docking lie far above QSAR-fits to random-noise-added data. This shows that much of the error in the docking results follows neighborhood behavior and can be reproduced by the QSAR model in the fingerprint descriptor space. While the particular sources of neighborhood error lie beyond the scope of this study, deviation from the pure-noise QSAR-fit line identifies its presence.

The small correlation values between docking scores and pIC_{50} raises the question of whether the range is significant enough for interpretation. Given that docking accuracy is perhaps more commonly assessed by enrichment-of-actives statistical measures, we addressed this issue by comparing

correlation to receiver operating characteristic (ROC) scores, which are calculated as the area under the true positive vs false positive curve. Unlike correlation calculations, enrichment-based statistics require setting an activity threshold. Using a 5 μM IC_{50} threshold to separate actives from inactives, ROC scores range from 0.53 (fail) to 0.80 (borderline-good) and rank the 32 docking models in this study very similarly to ranking by correlation, with a Spearman (rank-based) correlation between ROC-score and docking/pIC_{50} correlation of 0.96. Applying a 3 μM IC_{50} threshold, ROC scores rank-correlate with a coefficient of 0.94. This agreement of ranks is significant, especially considering that the Spearman correlation between the 5 μM and 3 μM threshold-defined ROC scores is only 0.87. The fact that alternate statistics confirm the relative accuracies of the docking models argues that the dynamic range in correlation is significant and interpretable, which is particularly impressive given the "minuscule" correlations for the 19 PIM1 docking models. The absence of an activity-threshold parameter in correlation measurements is also attractive.

Another striking feature of the plots in Figure 8 is the positive trend between QSAR-fits and data set accuracy. The crystal structures that give the best docking results give the best QSAR-fits. (For PIM1, none of the docking models produce scores with any significant correlation to experiment, so the trend remains untested in this case.) That better docking models generate scores more easily explained by 2D chemical descriptors suggests a potential for QSAR-fit to prospectively appraise various activity prediction models in the absence of experimental measurements. I.e. use the crystal structure whose docking results can best be fit by 2D-QSAR, i.e. that show the strongest neighborhood behavior, for the best estimate of IC_{50}.

QSAR-fits to these same docking scores were also calculated using alternative QSAR parameter combinations (Figure 9). While some QSAR parameters allow better fit to the docking data, the correspondence between better docking models (crystal structures) and better QSAR-fits holds irrespective of which 2D-QSAR parameters are used. Retaining more information capacity in the QSAR equation, either by including more fingerprint features or more latent variables in PLS, increases QSAR-fits for all docking models uniformly.

QSAR "Cleaning" Docking Scores. Given the resilience of the pIC_{50} signal in QSAR equations fit to substantially noisy random data (Figure 4), one wonders how much docking scores could be improved by QSAR smoothing. We tested error-filtering capacities using a host of QSAR parameter sets for the three docking data sets (Figure 10). For all docking score sets with at least a modicum of correlation to pIC_{50} ($R^2 > 0.1$), QSAR "cleaning" improves accuracy −and quite substantially for PDK1 PLS-regressed with one latent variable, with the largest improvement in pIC_{50} correlation from 0.11 to 0.43. This particular parameter set is not always optimal. For the best performing CSF1R crystal structure (docking accuracy: $R^2 = 0.26$), one latent variable performs worst of the tested parameter sets, whereas for the second best CSF1R docking model, one latent variable PLS performs best.

Ideally the best QSAR parameter choices could be selected prospectively, in other words, without pIC_{50} data for comparison. However, we have not identified any method for this selection. Cross-validating the reproduction of docking score does not help select QSAR models that predict pIC_{50} with better accuracy. Specifically, Q^2 calculated via a 5-fold split of the training data does not trend with the boost in QSAR-smoothed docking accuracy (results not shown). Certainly the most information-rich QSAR models, containing more chemical descriptors regressed using many PLS latent variables or by OLS regression, will better reproduce docking scores but tend to overfit, with limited cleaning power. More parsimonious QSAR models (e.g., few chemical descriptors modeled by few latent variables in PLS) are resistant to overfitting but still do not guarantee excellence.

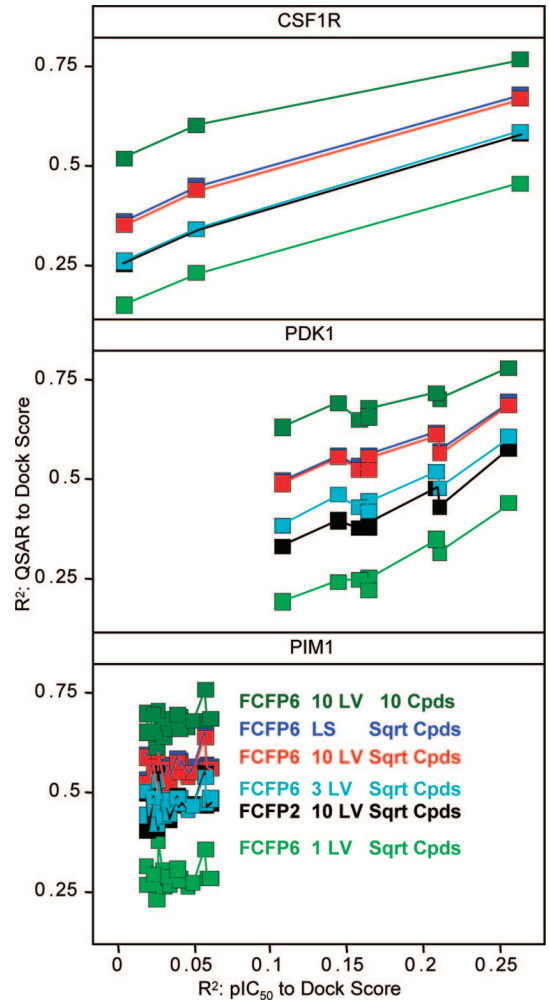

Figure 9. Docking: QSAR fit vs docking accuracy. Six sets of QSAR-fits to docking scores are plotted as functions of docking score correlation with experiment. Each connected set of symbols corresponds to a particular color-coded choice of QSAR parameters: either a two-bond diameter (FCFP2) or a six-bond diameter (FCFP6) fingerprint, either least-squares regression (LS) or PLS with 1, 3, or 10 latent variables (LV), and either 10 or a $\sqrt{N}$ feature-occurrence-threshold.

The minimum-threshold in docking score accuracy required to benefit from QSAR-fitting (Figure 10) parallels results by Klon et al.[20] That study showed a case where false-positives in the training data resulted in Bayesian models that degraded accuracy in the activity ranking compared to raw docking scores. Similarly in our study, regression on the worst CSF1R docking scores leads to a QSAR equation with negative correlation to pIC_{50} (points below the line in Figure 10).

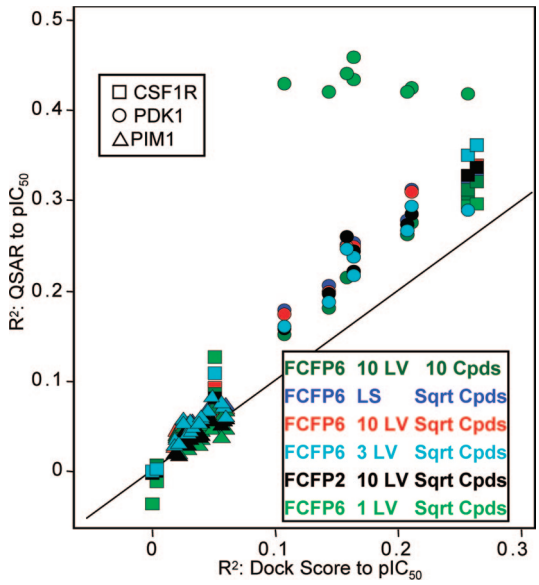

Figure 10. QSAR-cleaned docking accuracy vs docking accuracy. Docking scores from the three CSF1R, eight PDK1, and 19 PIM1 crystal structures, using each kinase's respective data set, were QSAR-cleaned using six different parameter selections, as summarized and color-coded by the legend. The docking score accuracy is plotted on the *x*-axis, with QSAR accuracy on the *y*-axis. The thin line indicates the $x = y$ diagonal. Negative values indicate negative correlation.

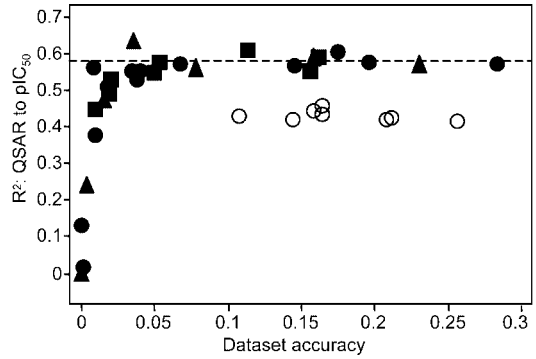

Figure 11. PDK1 one-latent-variable QSAR accuracy. QSAR-cleaned docking accuracies for PDK1 using one latent variable PLS equations (open circles), taken from Figure 10, are plotted amongst QSAR-cleaned accuracies of noise-added pIC_{50} (opaque symbols), likewise modeled with one latent variable. The neighborhood limit (QSAR-fit to pure pIC_{50}) is plotted as a dashed line.

The large accuracy boost for PDK1 docking score from PLS smoothing using one latent variable deserves special comment. Intriguingly, all docking score sets boost to nearly the same QSAR accuracy (0.42−0.46), despite a more substantial range in raw docking accuracy of 0.11−0.26. Examining the QSAR accuracies of noise-added pIC_{50} data sets plotted along with the docking-trained QSAR accuracy helps to explain the interplay between signal, neighborhood error, and non-neighborhood error (Figure 11). While the "neighborhood limit" for PDK1's pIC_{50} data set is much lower with one latent PLS variable (0.58) compared to, for example, 10 latent variables (0.79, Figure 8), the former's more underdetermined QSAR equation demonstrate remarkable resilience in the face of noise, with QSAR accuracy essentially constant (albeit with scatter) down to data set accuracy of 0.04. Even the three data sets with raw accuracy in the range 0.008−0.01 yield QSAR accuracies ranging from 0.38 to 0.56. At data set accuracy of 0.003, QSAR accuracy is still 0.24. Given the noise resilience with the one-latent-variable PLS QSAR model, the small range in docking-trained-QSAR accuracy is less surprising and suggests that these eight sets of docking scores differ primarily in how much non-neighborhood error they contain. In contrast, the neighborhood errors introduced by the docking procedure, which are in addition to the neighborhood error of applying QSAR to pIC_{50}, appear to be nearly equal among the eight sets.

Docking Scoring Function Selection. While a ligand design project may not have the luxury of selecting from multiple crystal structures, invariably multiple scoring functions implementing different principles must be considered, with the best choice depending on the docking system.[4] Could QSAR-fit help guide this choice? To answer this question, the Flo+ minimized docked poses were rescored with the three functions available in DockIt: the 'native' DockIt molecular mechanics score (DockIt MM), an empirical piecewise linear potential (PLP)[33] score, and Muegge and Martin's knowledge-based potential of mean force (PMF)[34] score. For each crystal structure, poses scoring greater than +40 on any of the three rescoring functions were eliminated as these likely reflect clashes upon imposing the new scoring functions. For the remaining poses, the best score among all poses for a given compound and crystal structure defines each scoring function's "score", with this operation performed independently for each of the four scoring functions.

Figure 12 plots docking accuracy against fits to 2D-QSAR models independently trained on the four docking score sets: Flo+, DockIt MM, PLP, and PMF. In all cases, choosing the scoring function with the best QSAR-fit would be the best or nearly best choice.

For PDK1, the trend is quite convincing that QSAR-fit correlates with docking accuracy (trend within each symbol shape in Figure 12, middle panel). For seven of eight crystal structures, the scoring function with the largest QSAR-fit is also the most accurate. In five of eight cases, the rank agreement is perfect. As seen above for Flo+, selecting a crystal structure for a particular scoring function is also assisted by QSAR-fit with the three additional scoring functions (trend within each symbol color in Figure 12).

For CSF1R, the best QSAR-fitting scoring function is also the most accurate in two of three crystal structures. The missed case is the poorest performing crystal structure. Interestingly for CSF1R, the QSAR-fit versus docking accuracy trend is much stronger across crystal structures for a single scoring function than across scoring functions on a given structure. This result hints that each scoring function must first be "normalized" for inherent fit-ability before comparing across structures. However, if such a propensity exists, it is not transferable, as PDK1 displays no apparent offset between scoring functions.

For PIM1, all crystal structures with all scoring functions perform poorly ($R^2 < 0.06$), so there is little dynamic range to test the hypothesis. Still, for 11 of 20 crystal structures, the best QSAR fitting scoring function is also the most accurate, whereas random would predict 5/20. In all cases, the PMF scoring function has the largest QSAR-fit for PIM1.

In order to test whether these additional scoring functions can benefit from QSAR-smoothing, the docking accuracy was plotted against the 2D-QSAR-smoothed accuracy (Figure 13).

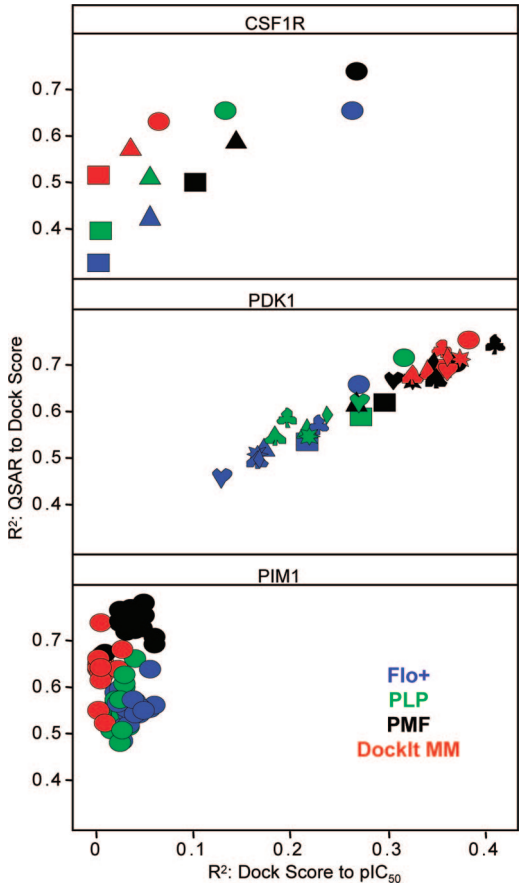

Figure 12. 2D-QSAR fit for multiple docking scoring functions. Docking accuracy is plotted against QSAR-fit for activity data sets corresponding to three kinases. Four scoring functions were tested, each color-coded. For CSF1R and PDK1, each symbol shape corresponds to a different crystal structure. For PIM1, results for the 19 crystal structures employ the same symbol shape.

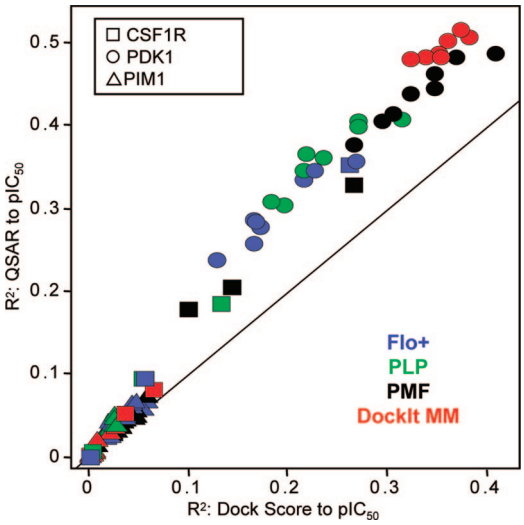

Figure 13. QSAR-cleaned docking accuracy with multiple docking scoring functions. Score sets from docking into the three CSF1R, eight PDK1, and 19 PIM1 crystal structures by four scoring functions (color-coded) were QSAR-cleaned using default QSAR parameter choices (see Methods). The docking score accuracy is plotted on the *x*-axis, with QSAR-cleaned docking score accuracy on the *y*-axis. Kinase identity is coded by symbol shape. The thin line indicates the $x = y$ diagonal.

Indeed, the trend seen for Flo+, that the 2D-QSAR cleaned predictions correlate better with experiment than the docking scores themselves, extends to all scoring functions. Plotting docking accuracy versus QSAR-cleaned docking accuracy produces a very linear relationship, irrespective of target, crystal structure, or scoring function. Using the "default" QSAR parameter choices, the improvements upon QSAR-smoothing depend primarily on the original docking accuracy. Excluding the docking models with insignificant accuracy, QSAR-smoothing improves accuracy by ~0.1 R^2 units.

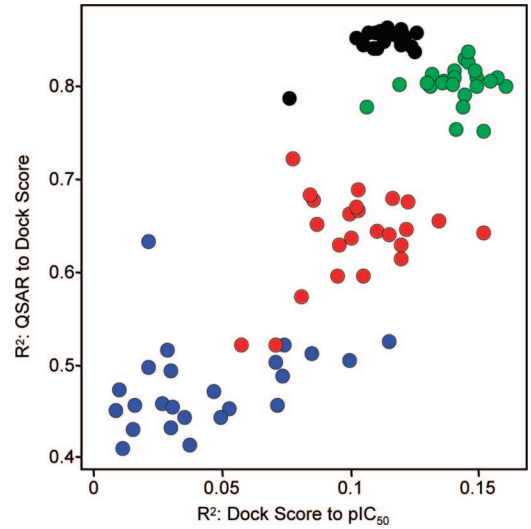

Figure 14. 2D-QSAR fit for HIV-1 protease docking. Four docking scoring functions (color coded as in Figure 13) were employed with 24 crystal structures. Docking accuracy is plotted against QSAR-fit.

HIV-1 Protease Docking Model Selection. The very intriguing result, that 2D-QSAR can help to select one of the best docking models from several, motivated analyzing an additional nonkinase system, HIV-1 protease. A public activity data set[23] was docked by DockIt into 24 PDB[26] crystal structures, minimized and scored by Flo+, and rescored by Dockit-MM, PLP, and PMF. 2D-QSAR models were then trained on each of the 96 sets of docking scores (Figure 14).

Despite a more limited docking-accuracy range, relative to PDK1 and CSF1R, a fairly convincing positive relationship between docking accuracy and QSAR-fit still emerges in Figure 14. R^2 for the 96 points in Figure 14 is 0.64, meaning QSAR-fit explains the majority of variance in docking accuracy. The *P*-value associated with this R^2 value is <0.0001.

Unlike PDK1 and CSF1R, scoring function appears to be the dominant factor in accuracy. Crystal structure choice plays a more limited role for HIV-1 protease docking. This probably reflects the greater conformational flexibility of kinases, rendering crystal structure selection a more critical task. PMF in particular shows a very tight docking accuracy range, independent of structure choice. Similar to CSF1R, the different scoring functions show some systematic off-diagonal deviation, revealing neighborhood error differences attributable exclusively to scoring function.

Training 2D Models from Target-Tailored Docking to Kinases. The improvements in prediction accuracy following QSAR regression on docking scores, where a general scoring function is used, can also extend to target-specific scoring functions. Surrogate AutoShim is a target-

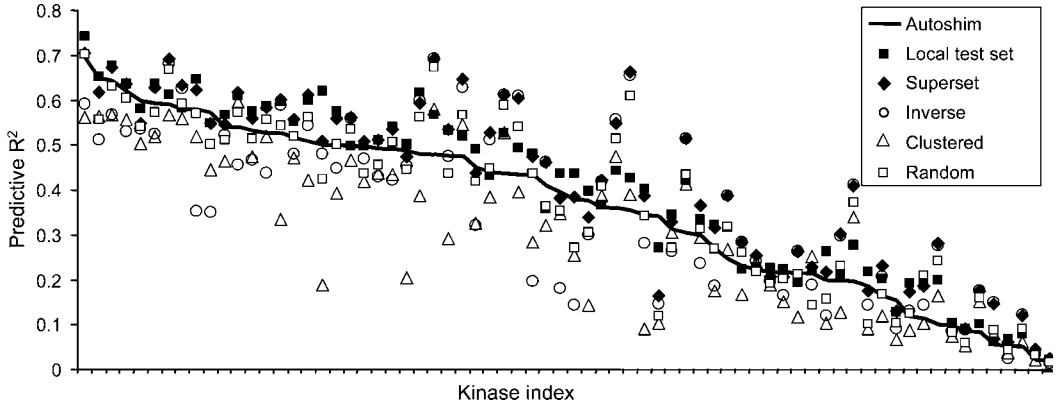

Figure 15. 2D-QSAR predictions trained on AutoShim docking results for 5 training sets. Surrogate target-tailored docking models built from docked poses by the AutoShim method[32] were evaluated against 25% withheld test sets for 70 kinases (solid line, no symbol). For each kinase, 2D-QSAR models were trained on AutoShim scores using five types of QSAR training sets (described in Methods and Figure 2). All were evaluated against the 25% withheld data available for that kinase.

tailored docking method that rescores docked ligand poses with a scoring function augmented with "shims", i.e. pharmacophoric-features, trained on experimental activity data, using a universal ensemble kinase model of 16 crystal structures.[32] AutoShim predictions were regressed against 2D chemical descriptors by PLS (Figure 15). All AutoShim predictions used for training 2D-QSAR models originate from a withheld 25% of the total available pIC_{50} data for a particular kinase; the other 75% were used to train the AutoShim models themselves and were not directly used in the 2D-QSARs (see Methods). The number of test compounds available to each 2D-QSAR (termed each kinase's "local" test set) ranges from 186 to 13,292. Twenty-two test sets have fewer than 250 compounds. Only 10 kinases have more than 2000 test compounds with assay data.

In 62 of 70 cases, QSAR fitting the local test set improves correlation with pIC_{50}, but the average change in R^2 is a comparatively modest +0.035. However, many of these AutoShim docking models already had high correlations with activity. Furthermore, even a 2D-QSAR model that performs as well as the much more expensive docking method is useful for economically searching large databases.

One important issue for practical application of these methods is the dependence on the 2D-QSAR training set. Ideally, useful 2D models could be learned from 3D models using a diverse and relatively small compound set and then applied to large compound sets in order to capitalize on the computational speed advantages of 2D-QSAR.[35]

To address compound set dependence, four additional 2D-QSAR training sets were assembled. Note that the 2D-QSARs are trained only on computational AutoShim docking results and require neither experimental data for the 2D-QSAR training set nor docking results for the 2D-QSAR prediction compounds. The new 2D-QSAR training sets added docking results for compounds which lacked experimental assay results, in addition to the withheld 25% of each individual kinase's local test set for which experimental data were available for evaluation. However, all 2D-QSAR equations were evaluated over their local test sets. For comparison, 2D-QSAR models trained on the 75% set of experimental pIC_{50} data used for training AutoShim were also evaluated over the 25% local test sets.

The most frequently successful method for improving predictive-R^2 was to both train 2D-QSAR on docking predictions and evaluate the QSAR model over the "local" set results, which improves R^2 in 62 of 70 cases (Table 1). Nearly as consistently, the "superset" trained 2D-QSAR models improve predictive-R^2 on the local test set in 57 cases. However, the magnitude of the average improvement in predictive R^2, relative to AutoShim alone, is slightly higher for "superset" (+0.054) versus "local" (+0.035), with some quite dramatic "superset" improvements. The "superset" vs raw AutoShim R^2 difference for kinases with small "local" test sets (22 cases < 250 compounds) averages an impressive +0.10, competitive with 2D-QSAR trained directly on the experimental data, which averages 0.11 R^2 units better than raw AutoShim over these kinases.

Table 1. 2D-QSAR Cleaning AutoShim Predictions

set	average R^2	no. of improved (of 70)[a]	average R^2 change[a]
AutoShim (no 2D)	0.362		
local	0.397	62	0.035
inverse	0.360	31	−0.002
superset	0.416	57	0.054
clustered	0.309	17	−0.053
random	0.372	34	0.010
2D-QSAR[b]	0.463	61	0.101

[a] Number improved and R^2 change is relative to AutoShim predictions without 2D-QSAR postprocessing. [b] 2D-QSARs were trained on pIC_{50} values of 75% training data and applied to the 25% withheld "local" test set. These "pure" 2D-QSAR results do not exemplify QSAR cleaning but are provided for reference.

Other metrics besides linear correlation likewise capture these improvements. For 18 of these 22 data sets with at least 5 hits (using 3 μM as a threshold), "superset" 2D-QSAR cleaning improves the area under the ROC curve from 0.85, on average, to 0.89, while the Spearman rank-based correlation improves from 0.45 to 0.54 on average. Therefore, supplementing with additional "superset" 3D-QSAR predictions appears to broaden the base of the 2D-QSAR models and improve accuracy.

Training on clustered data sets reduces the prediction accuracy for the local test set relative to AutoShim predictions (average difference: −0.053). 2D-QSAR models trained on the random set of the same size did not affect predictions (+0.01). The reduced performance of clustered relative to random probably originates from the activity distribution and not simply that fewer compounds train the models. The

clustering scheme used here, in which the compound set is successively divided until no cluster radius is larger than 0.6 Tanimoto distance without "rescuing" singletons, yields a very diverse compound set which reduces the fraction of active compounds. Alternative selection methods might better preserve active examples[36] and train more predictive 2D-QSAR models.

2D-QSAR models trained on the "inverse" sets, arguably the most ambitious test, because the 2D-QSAR is trained entirely on compounds lacking IC_{50} data, roughly retains the AutoShim accuracy; 31 models were improved by 2D-QSAR, while 39 decline.

In summary, applying 2D-QSAR to 3D-predictions works better in 'cleaning' mode, where the 2D-QSAR model is limited to rescoring the 3D-predictions used to train the 2D-QSAR, than in 'prediction' mode, where 2D-QSAR generates predictions for compounds without 3D-QSAR predictions. The modest accuracy improvement with "superset" training suggests that additional docking examples may further stabilize the activity function. Despite the fact that 2D-QSAR models trained from clustering-reduced data sets perform the poorest, the accuracy preservation in "inverse" trained QSAR models gives hope that 3D-structure based activity prediction models trained on internal data can be "learned" in 2D chemical structure based models and then applied to external compound sources if the training set is properly designed.

DISCUSSION

A primary appeal to restating 3D prediction schemes with a 2D model is to extend the scope of applicability by benefiting from the greater speed of 2D models. In this capacity, researchers at Arqule have advocated building QSAR models from docking over a limited and diverse compound set in order to prioritize expensive docking calculations over a larger compound set.[35] Prior to this, researchers at Novartis explored using 2D-QSAR to 'clean' noisy docking scores by building a QSAR model on docking scores and applying the model to the identical compound set, yielding an additional score for each compound.[20]

Given that 2D-QSAR trained directly on experimental data gives much better prediction accuracy than plain docking and slightly better even than target-tailored docking, why bother with expensive docking at all? Why not just use 2D-QSAR trained on activity data directly? 2D-QSAR models rely heavily on topological similarity. Conventional wisdom says that docking models, which do not "know about" 2D topology, work equally well (or poorly) on structurally dissimilar actives topologically unrelated to the training set. I.e., docking can make larger "scaffold hops", albeit at the cost of lower accuracy, leading to many false positives. There is little direct literature support for this generally held belief, and it is difficult to imagine a systematic study that would confirm it. However, a recent paper that compiled "examples of successful scaffold-hops" indirectly supports it.[37] Only 2 of the 21 examples in the survey used strictly 2D methods. Hopefully, by training a 2D-QSAR on docking predictions that include examples of these structurally novel actives predicted by docking (or AutoShim), the 2D-QSAR can be trained to recognize some of these distant active chemotypes and quickly retrieve them from large databases, with many of the false positives "smoothed away". Figure 16 presents

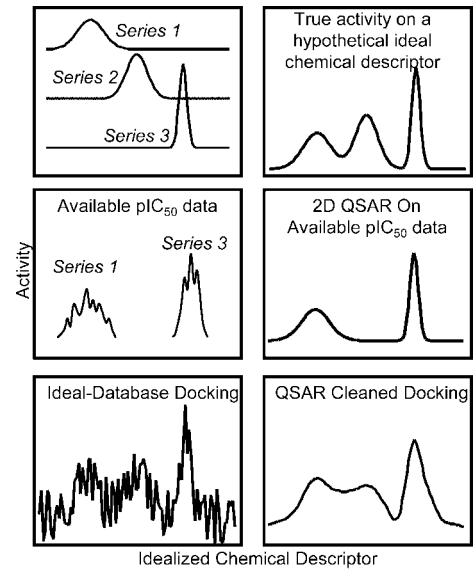

Figure 16. Hypothetical discovery of novel chemotypes. Three active chemotypes are posited, of which only series 1 and 3 have known examples with measured IC_{50}s. 2D-QSAR accurately predicts the activity of members of these series but misses the activity of series 2. Docking models make noisy predictions but do predict some active members of series 2. 2D-QSAR trained on these docking results now identifies the series 2 members in large virtual libraries or commercial databases.

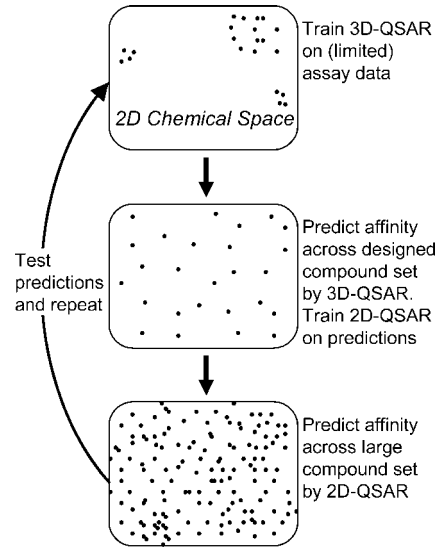

Figure 17. Workflow to boost experimental activity with 3D predictions in 2D-QSAR screening: (1) train a moderately predictive target-specific AutoShim scoring function based on data from 2 or 3 chemical series, (2) dock and predict the activity of a designed subset of readily available compounds or virtual libraries and train a 2D-QSAR on the docking results, (3) predict activity across the full compound set, and (4) order and test predicted actives. Add new data to the AutoShim training set and repeat.

a hypothetical situation with idealized results that illustrates the common drug design situation, where compounds with assay data do not explore enough chemical space to train a 2D QSAR that predicts actives with a novel scaffold that may exist in a larger compound collection. Training a 2D-QSAR on docking results that include these additional species might retrieve these additional actives. Figure 17 diagrams a workflow to boost a 2D-structurally homogeneous activity data set by training 3D methods such as AutoShim, in order to facilitate finding novel chemotypes such as "Series

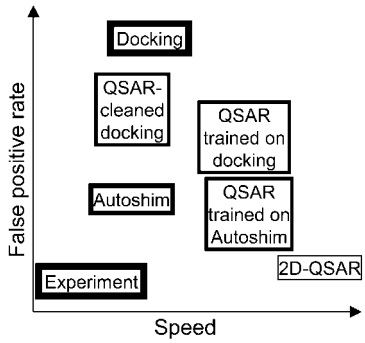

Figure 18. Speed, false-positive rate, and scaffold-hopping potential. Several methods are placed according to their relative speed and false-positive rates. The line-width of boxes indicates potential for scaffold hopping, with thicker lines corresponding to higher potential.

2" in Figure 16. Relative to 2D-QSAR, 3D-models offer a greater hope of finding new chemical scaffolds; however, noise and compute-time may hinder their effective use. QSAR-cleaning could potentially amplify the docking signal. Figure 18 diagrams the relative merits of these various methods, placing each according to speed, false positive rate, and scaffold hopping potential.

One outstanding question is whether both greater accuracy and speed-up can be simultaneously realized, in other words, can QSAR models trained on plain or AutoShim docking scores across a designed compound set yield greater predictive accuracy outside this training set when applied to large corporate compound collections or huge commercial vendor catalogues? Our results training 2D-QSAR models on AutoShim scores lend some insight. The 2D-QSAR trained on the "inverse" set, in which the evaluation compound set was neither used to train AutoShim, nor used to train the 2D-QSAR model, performs nearly identically to direct AutoShim docking. These results, of course, have the caveat that corporate compound collections, and particularly the subset with IC_{50} values, contain many congeneric medicinal chemistry series, so that test sets and training sets do not differ substantially in terms of chemical structure unless careful measures are taken (which were not). Perhaps the most pertinent message of Figure 15 is the importance of training set design. Restricting 2D-QSAR training to docking results for compounds lying in chemical cluster-centers produces the worst models. This might be a result of the sparse sampling of actives when extreme diversity is enforced. The most extensive "superset" performs best, perhaps due to the most opportunity to average out non-neighborhood error, as seen in Figure 5 with synthetic data. With a linear fitting procedure, random noise should decrease as $\sqrt{N}$. Additional research into training set design in this context is warranted. It will be interesting to see how well principles for screening library design[38–40] transfer to this problem.

In addition to practical prediction improvements with 2D-QSAR cleaning, the methods can also unveil error-structure in activity estimates. Comparing QSAR fits to actual pIC_{50} data, docking activity estimates, and pIC_{50} data with added error shows that QSAR-cleaning can only reduce random error and at the neighborhood-limiting cost of imperfect QSAR technology. The results in Figures 8, 9, 12, and 14 consistently show that the crystal structures and scoring functions for which docking scores can best be fit by 2D-QSAR also give the most reliable activity predictions. If neighborhood error were exclusively, in effect, replacing activity signal in the poorer docking models, then QSAR-fit would not distinguish good docking models from bad. Therefore, noise, not neighborhood error, must be overriding the activity signal in the poorer docking models. The PDK1 docking score set in the underdetermined one-latent-variable defined chemical space particularly exemplifies this empirical result (Figures 10 and 11). QSAR cleaning for all crystal structures boosts PDK1 docking accuracy to essentially the same QSAR accuracy, consistent with all dockings suffering from an identical neighborhood limit, reflecting imperfect QSAR fit to pIC_{50}, plus an identical neighborhood error imparted by docking. Essentially all non-neighborhood error, which varies in amount between docking score sets, is smoothed away by the 2D-QSAR.

CONCLUSIONS

We have shown that the 2D-QSAR fit to noisy data decreases linearly with random noise (Figure 3). However, instead of confounding 2D-QSAR, noise escapes fitting, and hence is largely smoothed away, along with some portion of the activity information (Figure 4). The strong QSAR-fits to generally low-accuracy docking scores indicate that docking introduces significant neighborhood errors that can be fit by chemical descriptors (Figure 8). In other words, docking systematically mis-scores entire chemical neighborhoods. Despite these substantial systematic errors in docking, QSAR-fits are highest to the better docking models (Figures 8, 9 and 12) suggesting that having less random error distinguishes better docking models from worse. Further, the non-neighborhood error in docking, either with a general scoring function or target-specific scoring function, can often be attenuated by 2D-QSAR (Figures 10, 13 and 15). These results suggest that a 2D-QSAR fit to docking scores over a well-designed compound set can make scaffold hops in large databases with less time and fewer false positives than either docking or direct 2D-QSAR (Figures 16 and 17).

ACKNOWLEDGMENT

The authors appreciate helpful comments from Scott Dixon and Meir Glick.

APPENDIX

The linear relationship between QSAR-fit and error can more easily be understood by considering the case of a single chemical descriptor, X, as opposed to the linear combination of descriptors that define the PLS-based QSAR equations. X correlates with activity, Y, as

$$R^2 = \left(\frac{\langle XY \rangle - \langle X \rangle \langle Y \rangle}{\sqrt{\langle X^2 \rangle - \langle X \rangle^2} \sqrt{\langle Y^2 \rangle - \langle Y \rangle^2}} \right)^2 \quad (1)$$

The scale and range invariance of correlation allows us, for simplicity, to normalize X and Y to have a mean of zero and unit-variance without altering our argument. Under these conditions, eq 1 simplifies to

$$R^2 = \langle XY \rangle^2 \quad (2)$$

The noise-added activity value, Y', can be generated by

combining a proportion, a, of the pure activity distribution, Y, to a proportion of zero-centered, unit-variance random error, e. This proportionality can vary from 0 (pure noise) to 1 (pure activity). Y' is likewise centered on zero with unit variance by defining Y' as:

$$Y' = Y\sqrt{a} + e\sqrt{1-a} \quad (3)$$

In Figure 3, the abscissa plots correlation between pure activity and noise-added activity, which corresponds to $\langle YY' \rangle^2$ in this example. Substituting in eq 3,

$$\langle YY' \rangle^2 = \langle Y(Y\sqrt{a} + e\sqrt{1-a}) \rangle^2 = \langle Y^2\sqrt{a} \rangle^2 + \langle Y^2\sqrt{a} \rangle \langle Ye\sqrt{1-a} \rangle + \langle Ye\sqrt{1-a} \rangle^2 \quad (4)$$

Random error, by definition, has zero correlation with all other independent distributions, so that all terms on the right of eq 3 fall out except $\langle Y^2\sqrt{a} \rangle^2$, which further reduces to $a\langle Y^2 \rangle^2$. Because Y has unit variance, the abscissa simply reports a.

The ordinate of Figure 3 plots correlation between chemical descriptors and noise-added activity, which corresponds to $\langle XY' \rangle^2$ in this example and similarly expands to:

$$\langle XY' \rangle^2 = \langle X(Y\sqrt{a} + e\sqrt{1-a}) \rangle^2 = \langle XY\sqrt{a} \rangle^2 + \langle XY\sqrt{a} \rangle \langle Xe\sqrt{1-a} \rangle + \langle Xe\sqrt{1-a} \rangle^2 \quad (5)$$

If error is independent of the chemical descriptor, then all terms with e in a product reduce to zero such that the ordinate exclusively reports $a\langle XY \rangle^2$; in other words, the correlation between pure activity and the chemical descriptor, modulated by the proportionality constant. Hence, Figure 3 is expected to plot chemical descriptor correlation to pure activity at $a = 1$, and cross through the origin at $a = 0$ with a linear relationship throughout. In our actual cases, the QSAR equation is not independent of the error given that the equation is fitted to the error-containing activity estimate, Y', and hence "chance correlations" enter the relationship. Reducing the number of latent variables reduces the amount of error that can "bleed" into the QSAR equation, which analogously has been discussed in depth for factor analysis in seminal work by Malinowski.[41] The Xe containing terms thus do not fall out in eq 5, resulting in the up shifting of the trend lines in Figure 3. In the ideal case of many data points, chance correlations, and the Xe term, would disappear.

REFERENCES AND NOTES

(1) 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.

(2) Warren, G. L.; Andrews, C. W.; Capelli, A. M.; Clarke, B.; Lalonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Peishoff, C. E.; Head, M. S. A Critical Assessment of Docking Programs and Scoring Functions. *J. Med. Chem.* **2006**, *49*, 5912–5931.

(3) 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.

(4) Perola, E.; Walters, W. P.; Charifson, P. S. A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance. *Proteins: Struct., Funct., Bioinf.* **2004**, *56*, 235–249.

(5) Ha, S.; Andreani, R.; Robbins, A.; Muegge, I. Evaluation of docking/scoring approaches: A comparative study based on MMP3 inhibitors. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 435–448.

(6) Bissantz, C.; Folkers, G.; Rognan, D. Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. *J. Med. Chem.* **2000**, *43*, 4759–4767.

(7) Good, A. C.; Cheney, D. L.; Sitkoff, D. F.; Tokarski, J. S.; Stouch, T. R.; Bassolino, D. A.; Krystek, S. R.; Li, Y.; Mason, J. S.; Perkins, T. D. Analysis and optimization of structure-based virtual screening protocols. 2. Examination of docked ligand orientation sampling methodology: mapping a pharmacophore for success. *J. Mol. Graphics Modell.* **2003**, *22*, 31–40.

(8) Ewing, T. J. A.; Kuntz, I. D. Critical evaluation of search algorithms for automated molecular docking and database screening. *J. Comput. Chem.* **1997**, *18*, 1175–1189.

(9) Ajay; Murcko, M. A. Computational methods to predict binding free energy in ligand-receptor complexes. *J. Med. Chem.* **1995**, *38*, 4953–4967.

(10) Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E. Neighborhood behavior: A useful concept for validation of "molecular diversity" descriptors. *J. Med. Chem.* **1996**, *39*, 3049–3059.

(11) Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do structurally similar molecules have similar biological activity. *J. Med. Chem.* **2002**, *45*, 4350–4358.

(12) Kubinyi, H. Similarity and dissimilarity: A medicinal chemist's view. *Perspect. Drug Discovery* **1998**, *9–11*, 225–252.

(13) Walters, W. P.; Goldman, B. B. Feature selection in quantitative structure-activity relationships. *Curr. Opin. Drug Discovery Dev.* **2005**, *8*, 329–333.

(14) Hansch, C.; Fujita, T. Rho-Sigma-Pi-Analysis. A method for the correlation of biological activity and chemical structure. *J. Am. Chem. Soc.* **1964**, *86*, 1616–1626.

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

(16) Bohacek, R. S.; McMartin, C.; Guida, W. C. The art and practice of structure-based drug design: A molecular modeling perspective. *Med. Res. Rev.* **1996**, *16*, 3–50.

(17) Stone, M. Cross-Validatory Choice and Assessment of Statistical Predictions. *J. Royal Stat. Soc. Ser. B* **1974**, *36*, 111–147.

(18) Geisser, S.; Eddy, W. F. A Predictive Approach to Model Selection. *J. Am. Stat. Assoc.* **1979**, *74*, 153–160.

(19) Glick, M.; Jenkins, J. L.; Nettles, J. H.; Hitchings, H.; Davies, J. W. Enrichment of high-throughput screening data with increasing levels of noise using support vector machines, recursive partitioning, and Laplacian-modified naive Bayesian classifiers. *J. Chem. Inf. Model.* **2006**, *46*, 193–200.

(20) Klon, A. E.; Glick, M.; Thoma, M.; Acklin, P.; Davies, J. W. Finding more needles in the haystack: A simple and efficient method for improving high-throughput docking results. *J. Med. Chem.* **2004**, *47*, 2743–2749.

(21) Krumrine, J. R.; Maynard, A. T.; Lerman, C. L. Statistical tools for virtual screening. *J. Med. Chem.* **2005**, *48*, 7477–7481.

(22) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893.

(23) Chen, X.; Lin, Y.; Liu, M.; Gilson, M. K. The Binding Database: data management and interface design. *Bioinformatics* **2002**, *18*, 130–139.

(24) Wold, H. Estimation of principal components and related models by iterative least squares. In *Multivariate Analysis*; Krishnaiah, P. R., Ed.; Academic Press: New York, U.S.A., 1966; pp 391–420.

(25) *Pipeline Pilot*, *version 6.0*; Scitegic: San Diego, CA, 2006.

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

(27) *DockIt*, *version 1.5*; Metaphorics LLC: Santa Fe, NM, 2001.

(28) *Magnet*, *version 1*; Metaphorics LLC: Santa Fe, NM, 2001.

(29) Vieth, M.; Higgs, R. E.; Robertson, D. H.; Shapiro, M.; Gragg, E. A.; Hemmerle, H. Kinomics-structural biology and chemogenomics of kinase inhibitors and targets. *Biochim. Biophys. Acta, Proteins Proteomics* **2004**, *1697*, 243–257.

(30) ter Haar, E.; Walters, W. P.; Pazhanisamy, S.; Taslimi, P.; Pierce, A. C.; Bemis, G. W.; Salituro, F. G.; Harbeson, S. L. Kinase chemogenomics: Targeting the human kinome for target validation and drug discovery. *Mini-Rev. Med. Chem.* **2004**, *4*, 235–253.

(31) McMartin, C.; Bohacek, R. S. QXP: Powerful, rapid computer algorithms for structure-based drug design. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 333–344.

(32) Martin, E. J.; Sullivan, D. C. Pre-docking into a Universal Ensemble Kinase Receptor for 3D activity prediction, very quickly, without a crystal structure *J. Chem. Inf. Model*. **2008**, accepted for publication.

(33) Gehlhaar, D. K.; Verkhivker, G. M.; Rejto, P. A.; Sherman, C. J.; Fogel, D. B.; Fogel, L. J.; Freer, S. T. Molecular recognition of the inhibitor AG-1343 by HIV-1 protease: conformationally flexible docking by evolutionary programming. *Chem. Biol.* **1995**, *2*, 317–324.

(34) Muegge, I.; Martin, Y. C. A general and fast scoring function for protein-ligand interactions: a simplified potential approach. *J. Med. Chem.* **1999**, *42*, 791–804.

(35) Yoon, S.; Smellie, A.; Hartsough, D.; Filikov, A. Surrogate docking: structure-based virtual screening at high throughput speed. *J. Comput.-*

Aided Mol. Des. **2005**, *19*, 483–497.

(36) Taylor, R. Simulation Analysis of Experimental Design Strategies for Screening Random Compounds as Potential New Drugs and Agrochemicals. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 59–67.

(37) Schneider, G.; Schneider, P.; Renner, S. Scaffold-Hopping: How Far Can You Jump. *QSAR Comb. Sci.* **2006**, *25*, 1162–1171.

(38) Teig, S. L. Point - Informative libraries are more useful than diverse ones. *J. Biomol. Screening* **1998**, *3*, 85–88.

(39) Martin, E. J.; Critchlow, R. E. Beyond mere diversity: Tailoring combinatorial libraries for drug discovery. *J. Comb. Chem.* **1999**, *1*, 32–45.

(40) Miller, J. L. Recent developments in focused library design: Targeting gene-families. *Curr. Top. Med. Chem.* **2006**, *6*, 19-29.

(41) Malinowski, E. R. Theory of error in factor analysis. *Anal. Chem.* **1977**, *49*, 606–612.

CI700439Z