JOURNAL OF CHEMICAL INFORMATION AND MODELING

ARTICLE

pubs.acs.org/jcim

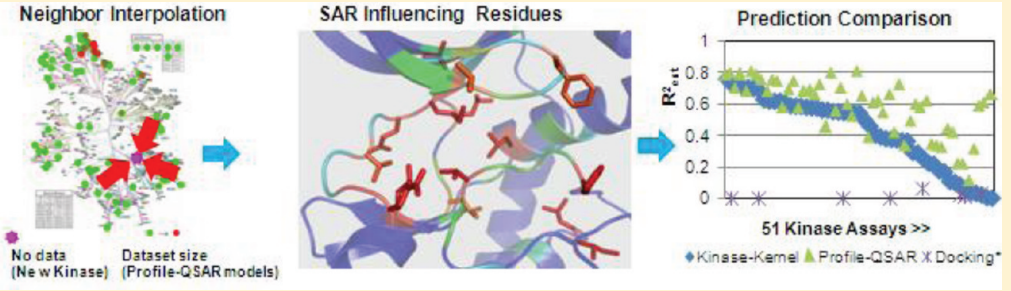
Kinase-Kernel Models: Accurate In silico Screening of 4 Million Compounds Across the Entire Human Kinome

Eric Martin* and Prasenjit Mukherjee

Oncology and Exploratory Chemistry, Global Discovery Chemistry, Novartis Institutes for Biomedical Research, 4560 Horton Street, Emeryville, California 94608, United States

Supporting Information

ABSTRACT:



Reliable in silico prediction methods promise many advantages over experimental high-throughput screening (HTS): vastly lower time and cost, affinity magnitude estimates, no requirement for a physical sample, and a knowledge-driven exploration of chemical space. For the specific case of kinases, given several hundred experimental IC_{50} training measurements, the empirically parametrized profile-quantitative structure−activity relationship (profile-QSAR) and surrogate AutoShim methods developed at Novartis can predict IC_{50} with a reliability approaching experimental HTS. However, in the absence of training data, prediction is much harder. The most common a priori prediction method is docking, which suffers from many limitations: It requires a protein structure, is slow, and cannot predict affinity.[1] Highly accurate profile-QSAR[2] models have now been built for roughly 100 kinases covering most of the kinome. Analyzing correlations among neighboring kinases shows that near neighbors share a high degree of SAR similarity. The novel chemogenomic kinase-kernel method reported here predicts activity for new kinases as a weighted average of predicted activities from profile-QSAR models for nearby neighbor kinases. Three different factors for weighting the neighbors were evaluated: binding site sequence identity to the kinase neighbors, similarity of the training set for each neighbor model to the compound being predicted, and accuracy of each neighbor model. Binding site sequence identity was by far most important, followed by chemical similarity. Model quality had almost no relevance. The median $R^2 = 0.55$ for kinase-kernel interpolations on 25% of the data of each set held out from method optimization for 51 kinase assays, approached the accuracy of median $R^2 = 0.61$ for the trained profile-QSAR predictions on the same held out 25% data of each set, far faster and far more accurate than docking. Validation on the full data sets from 18 additional kinase assays not part of method optimization studies also showed strong performance with median $R^2 = 0.48$. Genetic algorithm optimization of the binding site residues used to compute binding site sequence identity identified 16 privileged residues from a larger set of 46. These 16 are consistent with the kinase selectivity literature and structural biology, further supporting the scientific validity of the approach. A priori kinase-kernel predictions for 4 million compounds were interpolated from 51 existing profile-QSAR models for the remaining >400 novel kinases, totaling 2 billion activity predictions covering the entire kinome. The method has been successfully applied in two therapeutic projects to generate predictions and select compounds for activity testing.

INTRODUCTION

The fully loaded cost of an industrial high-throughput screen (HTS) on a typical collection of 1.5 million compounds can take 6−9 months and cost up to a million dollars.[3] Cherry picking a significant fraction of the collection, or ordering large numbers of compounds from external vendors, is likewise very expensive. The assays themselves are typically single concentration percent inhibitions, which makes hit-list triaging largely guesswork, and confirmation dose−response assays often result in high false positive and false negative rates. Reliable in silico methods are wanted to screen larger areas of chemical space and allow for medium-throughput screening (MTS) of smaller compound subsets directly in multiconcentration enzymatic and cellular screens, as well as for rational triaging filters to rescue false negatives and prioritize follow-up from expensive experimental HTS. Such a method should be fast, automated, have accuracy

Received: July 8, 2011
Published: December 01, 2011

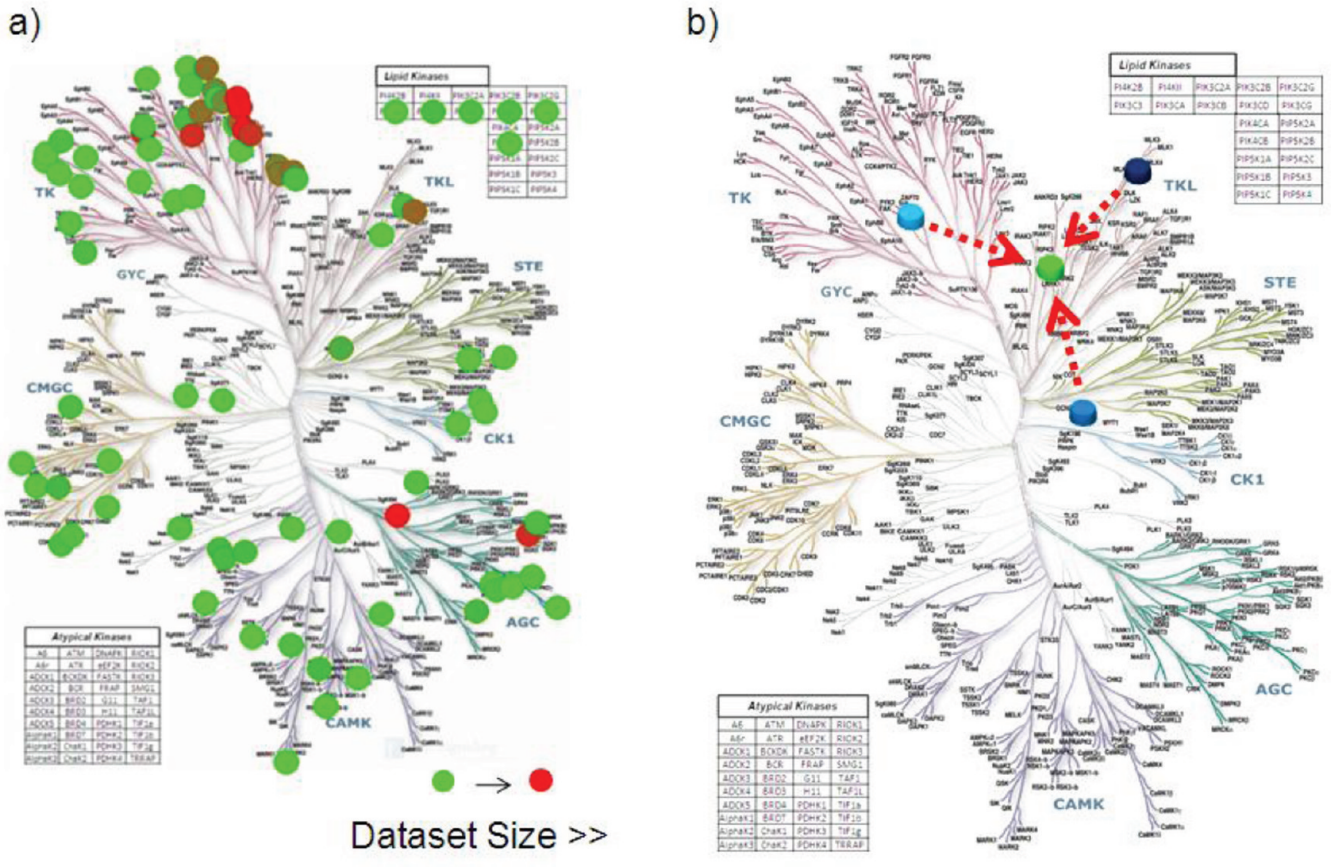


Figure 1. (a) A representation of the Sugen kinome tree showing the distribution of the 92 kinases with profile-QSAR models. The color indicates the number of experimental IC_{50}s. (b) An illustration of interpolation in the biological space. The green disk is a hypothetical novel kinase with no training data, while the blue discs represent three hypothetical neighboring kinases in the biological space having profile-QSAR models. The red arrows represent the process of interpolation. The activity of a compound against this new kinase is predicted as a weighted average of prediction of the same compound from the neighbor's profile-QSAR models.

similar to HTS, and estimate affinity rather than a mere yes/no activity.

Novartis's surrogate AutoShim[3,4] and profile-QSAR methods are two experimentally parametrized, protein family-based activity prediction techniques developed as in silico alternatives to HTS. The 3D surrogate AutoShim method uses several hundred IC_{50}s from an MTS IC_{50} screen to "shim" an ensemble of surrogate X-ray crystal structures, creating a target tailored scoring function that accurately predicts affinity. The same surrogate ensemble is used for an entire protein family, i.e. kinases, so the entire corporate archive is predocked just one time. The database of stored poses is retrained and rescored for new targets in the family, delivering almost instantaneous pose and activity predictions for each new kinase. Profile-QSAR is a 2D substructure-based meta-QSAR method. It uses activity predictions from a large database of initial Bayesian QSAR models from targets within a given protein family as chemical descriptors for a meta-QSAR trained to reproduce activity data for a new family member of interest. Together, these complementary 2D and 3D methods generate rapid, reliable, activity predictions from a modicum of affinity data. They are now routinely applied for kinase enzymatic and cellular activity and selectivity prediction, both for virtual screening and HTS triaging.

However, these experimentally parametrized models do typically require roughly 400−600 high-quality IC_{50} or EC_{50} measurements, with a range of affinity and structural diversity, for training these biochemical or cellular activity models. While profile-QSAR models have been developed for 92 kinases (115 kinase assays), this still leaves over 400 potential targets or antitargets from the human kinome where these techniques cannot yet be applied without developing a reliable assay and performing a substantial initial MTS and subsequent confirmation IC_{50}s, to obtain the requisite data to train the profile-QSAR and AutoShim models. An accurate a priori prediction method could "prime the pump", obviating the initial MTS and directly predicting the candidates for confirmation IC_{50} measurements. It could also be used in lieu of experiments for rough profiling.

An a priori activity prediction method can provide early chemical matter for early stage projects in target identification and validation or in the very early stages of hit discovery, before resources are justified for an expensive HTS. In addition, activity estimates for antitargets can be very important but rarely justify a HTS. An a priori method is valuable where expression of the target protein is difficult, the reagents being used for the assay are expensive, or the assay protocol itself is not amenable for conversion to HTS format. Furthermore, a kinomic[5,6] prediction method that extends prediction capability to the entire kinome

can be utilized for virtual chemokinomic profiling,[7] fishing for related targets,[8] or the identification of potential off-targets from the unchartered kinome space. Docking is by far the most common a priori virtual screening approach that does not require experimental parametrization. However, docking requires a protein structure, is time-consuming, and generally does not correlate with biological activity.[1] Systems biology[9−11] approaches generate networks between related targets based on data obtained by text mining literature sources for biological links or by correlating intersected data sets of compounds profiled on multiple targets. However, the confidence of the edges between the nodes is not always high due to conflicting literature results or statistically insignificant data sets being used to assign the edges.

Kernel methods[12−14] are a class of pattern recognition algorithms which identify relationships between points, vectors, sequences, text etc., using methods such as clustering, ranking, or correlations. Multivariate or spatial interpolations are kernel methods where the function to be interpolated is known at specific reference points, and the goal is to find its value at new points. Bitmap resampling is an example of 2D multivariate interpolation used in image processing.[15] A low-resolution picture is enhanced by superimposing a higher resolution grid on the original. The color codes for each pixel in the new grid are determined by interpolating from the color codes of the near neighbor pixels from the low-resolution original. Similarly, one can conceive the known SAR around the human kinome as a low-resolution picture made up of the reference set of previously studied kinases (Figure 1a), which already had data to build predictive profile-QSAR models (see Supporting Information). Kernel interpolation constructs a high-resolution picture of the kinome that contains all >500 kinases, where the affinity of a compound for each of the new kinases is assigned as a weighted average of the predicted affinities of neighboring kinases which already have existing models. In the absence of experimental SAR correlation between the reference kinases and the new kinases, binding site sequence identity can substitute as a measure of proximity in "cross-reactivity space" to identify and weight the near neighbors. Figure 1b illustrates an interpolation of a novel kinase with no training data (in green) from three neighbor kinases with models (in various shades of blue). Predictions for the novel kinase of interest are interpolated from profile-QSAR predictions for these reference kinases. This is uniquely possible for the kinase family due to the unprecedented accuracy of profile-QSAR models covering the entire kinome. The following sections describe the design, validation, and optimization of a formalism to carry out this interpolation that extends prediction capability to the whole kinome.

METHODS

Software and Resources. All calculations were carried out using either Pipeline Pilot v 8.0 (Accelrys, San Diego, CA) running on a 64-bit windows server, or using Python 2.4 and shell scripts on a 160 core Linux cluster with Intel Xeon processors (dual quad core blades at 2.9 GHz and 24GB of shared memory) running Red Hat 5.2. Spotfire DecisonSite 9.1 (Tibco, Palo Alto, CA) and Tibco Spotfire v 3.1 (Tibco, Palo Alto, CA) were used for visualization and analysis.

Protocols. The quality of the kernel models was quantified as the median R^2 between predicted and experimental IC_{50}s for 51 kinase assay data sets from the Novartis proprietary archive, ranging in size from 668 to 59 724 compounds. The compounds were

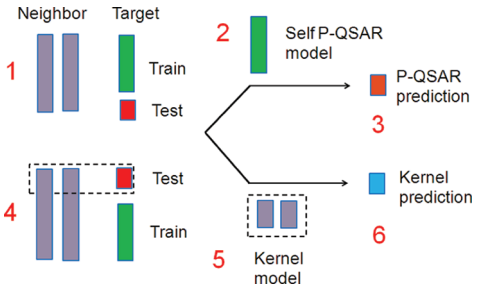


Figure 2. Study design for evaluating different kinase-kernel interpolation methods and comparing them directly to the trained profile-QSAR prediction method.

very diverse, some coming from broad profiling or hit-finding studies and a large portion coming from lead-optimization projects, where selectivity was emphasized. Apart from the 51 kinase assay data sets used for parametrization, two separate sets of 27 and 35 kinase assays were also evaluated as part of external validation. Empirically parametrized profile-QSAR models were used as a "gold standard" benchmark for accurate activity prediction. Binding site sequence identity, the "distance" measure in "kinase space", depends on the 2D alignment and selection of binding site residues. Initial studies used three kinase binding site sequence subset alignments: one generated internally, and two others generated by collaborators at Eidogen (Eidogen-Sertanty, Oceanside, CA). Choice of residues was subsequently further optimized with a genetic algorithm (GA). The GA was coded in Python, using the Pyevolve[16] library, with shell wrappers for data handling and parallel execution across the cluster. Ligand similarity was computed as Tanimoto similarity (T_c) between FCFP_6 fingerprints in Pipeline Pilot. Combination and preparation of the input data were carried out in Pipeline Pilot, then fed to Python interpolation scripts, and run in parallel across the cluster using shell wrappers. A separate Python script used NumPy[17] to combine the 3 individual alignments to generate the 46-residue align superset.

Partitioning of Training and Test Data Sets. The a priori kinase-kernel method does not require training data for the kinase of interest, but the empirically parametrized profile-QSAR models do. In order to compare their relative performance, a 25% test set was removed in both workflows, as shown in Figure 2. The upper row illustrates standard profile-QSAR modeling. In step 1, the complete set of IC_{50}s for each kinase (purple) is divided 75% for model training (green) and 25% for model testing (red). External test set predictions (orange) from the model trained in Step 2 are compared to experiment (Step 3) to obtain R^2_{ext} as a measure of model quality for each kinase. The workflow for kinase-kernel model predictions, illustrated in the lower row, also begins by extracting the same 25% test set. The kinase-kernel model is not trained, so it could be tested on the entire data set, but using the same test set as the profile-QSAR model permits a fair comparison. In the kinase-kernel scheme, neighbors are identified for a given target kinase in step 4, the neighbor's profile-QSAR predictions are made for the compounds in the target kinase's 25% test set (same test set as profile-QSAR) (step 5), and the kinase-kernel prediction (cyan) is interpolated from these predicted activities in step 6. The effectiveness of different neighbor selection and the interpolation schemes are compared as the median R^2 between predictions and experiment for the 51 kinases. The best kinase-kernel model is compared to the trained

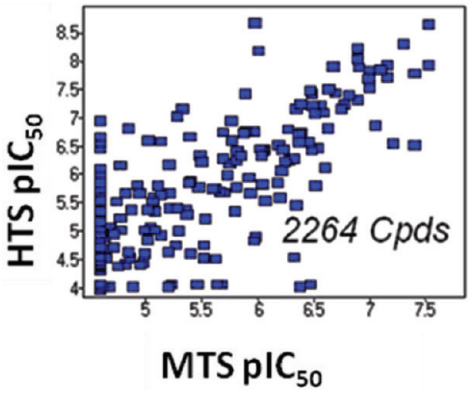


Figure 3. A plot of pIC_{50} values for 2264 compounds tested against PDK1 in two assay formats. HTS is a high-throughput, four concentration, log dilution IC_{50} assay. MTS is a careful, medium-throughput, eight concentration, half-log dilution IC_{50} assay.

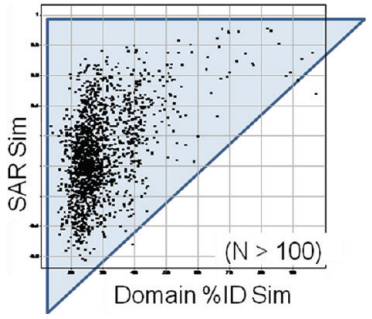


Figure 4. Each point represents a pair of kinases that have measured IC_{50}s for at least 100 common compounds. The *Y*-axis is SAR similarity (*R* between vectors of IC_{50}s for the common compounds between the kinase pairs). Whole kinase domain % sequence identity on the *X*-axis.

profile-QSAR method by the same metric. The kinase assay data sets used for the "external" sets of 17 and 35 validation assays that were not used in method optimization (see below) were also divided in the same way if profile-QSAR models were available. Where profile-QSAR models were not available for comparison, the full assay data sets were used for the kinase-kernel evaluation.

RESULTS AND DISCUSSION

Accuracy Expectations for an In silico HTS Alternative. The ideal goal would be an estimate of IC_{50} that correlates well with careful experiments, not just yes/no activity. Figure 3 shows a correlation of $R^2 = 0.6$ for phosphoinositide-dependent kinase-1 (PDK1), for 2264 compounds between a careful, medium-throughput, 8 concentration, half-log dilution IC_{50} assay and a high-throughput, 4 concentration, log dilution assay. A similar correlation between predicted and careful experimental IC_{50}s will be used as a benchmark for in silico methods to be at least approaching the reliability of a high-throughput IC_{50} experiment. By comparison,[1] correlations of IC_{50} with docking scores for diverse compounds rarely exceed $R^2 = 0.1$.

SAR Similarity vs Whole Kinase Domain Sequence Identity. The plot in Figure 4 of SAR similarity (IC_{50} correlation between kinase pairs with at least 100 compounds in common) vs whole kinase domain sequence identity (sequence percentage identity between pairs of whole kinase domains) shows a poor correlation

Table 1. Examples of Different Distance Attenuation Schemes[a]

	B_1	B_2	B_3	B_4	B_5	ΣB_i
Seq.Sim.	0.7	0.5	0.4	0.3	0.2	2.1
Seq.Sim(scaled)	0.33	0.24	0.19	0.14	0.1	1
$Seq.Sim^2$	0.49	0.25	0.16	0.09	0.04	1.03
$Seq.Sim^2$(scaled)	0.48	0.24	0.16	0.09	0.04	1
$Seq.Sim^{0.5}$	0.84	0.71	0.63	0.55	0.45	3.17
$Seq.Sim^{0.5}$(scaled)	0.26	0.22	0.2	0.17	0.14	1

[a] The first two lines show a simple linear attenuation based on binding site sequence identity. The third and fourth lines describe a square attenuation scheme, while the fifth and sixth lines show a square-root attenuation scheme.

($R^2 = 0.22$) but does show local similarity, i.e., the points lie in a triangular area above the diagonal. The filled upper triangle shows that kinases from distal branches (low sequence identity) can correspond to either high or low SAR similarity. However, the empty lower triangle indicates that high sequence identity does imply high SAR similarity. Local correspondence is enough, since the kinase-kernel interpolation only requires similar SAR among near neighbors. Given local kinase SAR similarity, and highly predictive profile-QSAR models covering the kinome, what remains is to optimize the interpolation weighting scheme.

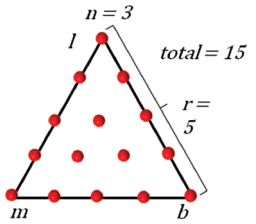
Optimizing the Kinase-Kernel Interpolation Weighting Scheme. Three primary factors described by eq 1 were considered to determine the weight from each profile-QSAR model to the kinase-kernel prediction. P_j, is the kinase-kernel activity prediction for compound *j* against a novel kinase. P_{ij} is the profile-QSAR activity prediction for the same compound *j* from one of a set of *n* neighboring reference kinases. Three neighbor model properties, B_i, M_i, and L_{ij}, determine the relative weights of the neighborhood models: binding site sequence identity, B_i, between the target and each neighbor kinase, "model quality", M_i, of the neighboring profile-QSAR models measured as 5-fold leave-group-out (LGO) Q^2, and "ligand similarity", L_{ij}, the average Tanimoto similarity of compound *j* with the five nearest compounds from the training set of model *i*. The three coefficients, *b*, *m*, and *l*, were optimized to adjust the relative impact of these three biases, subject to the constraint that $b + m + l = 1$ (see below), and *k* is the attenuation factor for scaling relative contributions of neighbor kinase profile-QSAR models as a function of binding site sequence identity.

$$P_j = b\sum_{i=0}^{n} B_i^k P_{ij} + m\sum_{i=0}^{n} M_i^k P_{ij} + l\sum_{i=0}^{n} L_i^k P_{ij} \quad (1)$$

Where P_j is the pIC_{50} prediction for compound *j* against the new kinase; B_i is binding site sequence identity of the neighbor kinase *i* to the new kinase; M_i is the model quality of the neighbor kinase *i* measured as 5-fold LGO Q^2; L_{ij} is the average Tanimoto similarity of ligand *j* with the five closest compounds in the training set of neighbor kinase *i*; P_{ij} is the pIC_{50} prediction from the profile-QSAR model of neighbor kinase *i* for compound *j*; *b*, *m*, and *l* are the individual coefficients with the constraint $b + m + l = 1$; and *k* is the attenuation factor used for decreasing the neighbor contributions by relative binding site sequence identity.

While the kinase-kernel model does not require training data for each new kinase of interest, the parameters involved in weighting the relative contributions of the neighbor models were empirically optimized to produce best predictions across 51 kinases

dx.doi.org/10.1021/ci200314j |J. Chem. Inf. Model. 2012, 52, 156–170



$$b + m + l = 1$$

$$total = \frac{(n + r - 2)!}{(n - 1)!\,(r - 1)!}$$

Figure 5. Illustration showing a simplified example of a mixture design using 3 variables ($n = 3$) b, m, and l and a step size of 0.25 ($r = 5$). Each point on the triangular response surface represents a specific combination of the 3 variables which follows the condition that $b + m + l = 1$.

with known IC_{50} data. The prediction step is computationally inexpensive. Predictions for the 51 assay data sets take only ~1 h, so the response surface could be determined in detail.

In addition to the weights for the three neighbor model properties, b, m, and l, two other kinase-kernel model parameters were required: the number of neighbor models to average, n, and how to attenuate the neighbor weight with distance, k. Three to eight neighbors were considered. Five functions were tested to attenuate the contributions of the near neighbor profile-QSAR models with distance, as exemplified in Table 1. Simplest was linear scaling, where the weighting coefficients are just the n sequence identities scaled to sum to 1. In squared polynomial scaling (squaring the sequence identities and scaling them to sum to 1), the relative contribution from more distant neighbors falls off more quickly. Conversely, square-root polynomial scaling (taking square root of the sequence identities and scaling to sum to 1) falls off less quickly than linear scaling. Cubic and cube-root polynomial scaling are analogous.

Binding Site Definition. Whole kinase domain sequence identity reflects evolutionary relationships, which need not relate to cross-reactivity. Binding-site sequence identity, which is more germane, requires a choice of binding site residues, and an alignment of those residues. For the initial determination of weights b, m, and l, the number of neighbor models to average and the distance-attenuation function, three pre-existing binding site definitions/alignments were tested: "A1" was balanced between residues from both the ATP pocket and the back pocket, "A2" was biased for ATP pocket residues, and "A3" was biased for back pocket residues. Binding site sequence identity was determined to be the most important weighting criterion, so the selection of binding site residues was further refined using a GA (see below).

Primary Interpolation Weights Optimized by Mixture Design. The relative weights of the three biases, b, m, and l, were optimized by a triangular mixture design,[18] as illustrated in Figure 5, which has the property that every point is normalized, i.e., $b + m + l = 1$. The three corners are the pure biases by a single factor, e.g., $b = 1$, $m = 1$ or $l = 1$. Other points of the response surface correspond to a specific combination of the three weighting coefficients. The step size was 0.01, giving 5151 combinations in each mixture design; 135 such mixture designs were run corresponding to 9 neighbor counts (2−10) by 3 binding site definitions and by 5 binding site distance attenuation functions. Each point in each design involved building kinase-kernel models for 51 kinases. The quality (median R^2 for 51 kinases) and b, m, l coefficients of the best model from each of the 135 mixture designs were analyzed to choose the best interpolation model.

Each bin in the histograms in Figure 6 shows results for each of the 5 attenuation functions for the best single weighting

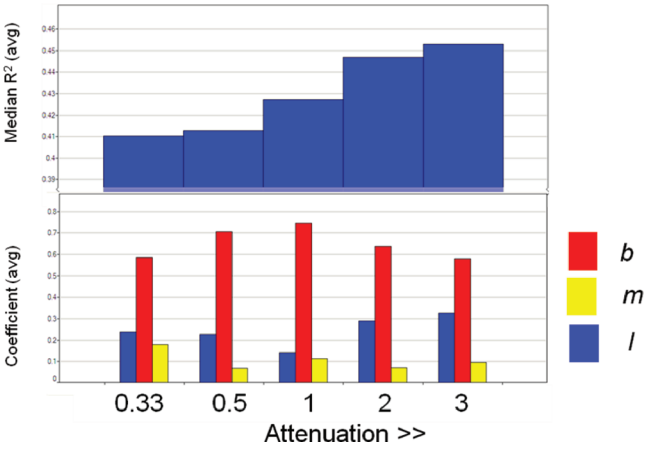


Figure 6. The top histogram shows performance (average of the median R^2 from the best performing combinations from 27 runs) of the 135 runs binned by binding site distance attenuation function. Each bin in the bottom histogram has three bars representing the average optimized coefficient value for the binding site sequence identity (b, in red), ligand similarity (l, in blue), and model quality, (m, in yellow) for the best performing combinations from 27 runs.

(choice of b, m, and l) averaged over 27 mixture designs (9 neighbor counts [2−10] by 3 binding site definitions). In the top histogram, which shows the average median R^2 for the 51 kinase models, performance progressively improves going from cube-root (0.33) to cubic (3) attenuation, showing that heavily weighting the nearest neighbors yields the best predictions. The histogram on the bottom breaks out the average of the three coefficients: b, m and l. Binding-site sequence identity is consistently the most important factor, but in these preliminary studies, ligand similarity (to the training set) appears to contribute to accurate predictions, especially with the nonlinear attenuation functions. Prediction accuracy is relatively insensitive to model quality. However, ligand similarity will also prove unimportant in the final, fully optimized models (see below).

Figure 7 shows that the optimal number of neighbors to include in the kinase-kernel was more ambiguous, and varied for different binding site sequence definitions. The X-axes vary the number of neighbor profile-QSAR models averaged for each prediction. The Y-axes are median R^2 over 51 kinase models for the best performing combination of b, m, and l coefficients from a single mixture design, for that combination of kinase neighbor count and binding site definition. Note that simply choosing the IC_{50} of the single nearest kinase neighbor as a surrogate for the new kinases gives a median $R^2 = 0.37$. The fully optimized kinase-kernel model will give a median $R^2 = 0.55$ (see below), showing the advantage gained by an optimized weighted average. Each plot has six series: the "gold standard" empirically parametrized profile-QSAR in black as reference and the series for the five distance attenuation functions. The cubic attenuation scheme (kernel_3) again is consistently the best performer. As expected, since profile-QSAR predictions are the actual kinase-kernel inputs, they out-perform the kinase-kernel predictions. The very best kinase-kernel prediction performance, with median $R^2 = 0.51$, used the ATP site-biased (A2) alignment, averaging predictions from 3 nearest neighbors. That is approaching profile-QSAR itself at 0.61. However, using three neighbors gave relatively poor results for the other binding site definitions. For cubic attenuation, the performance plateaued beyond about seven

dx.doi.org/10.1021/ci200314j |J. Chem. Inf. Model. 2012, 52, 156–170

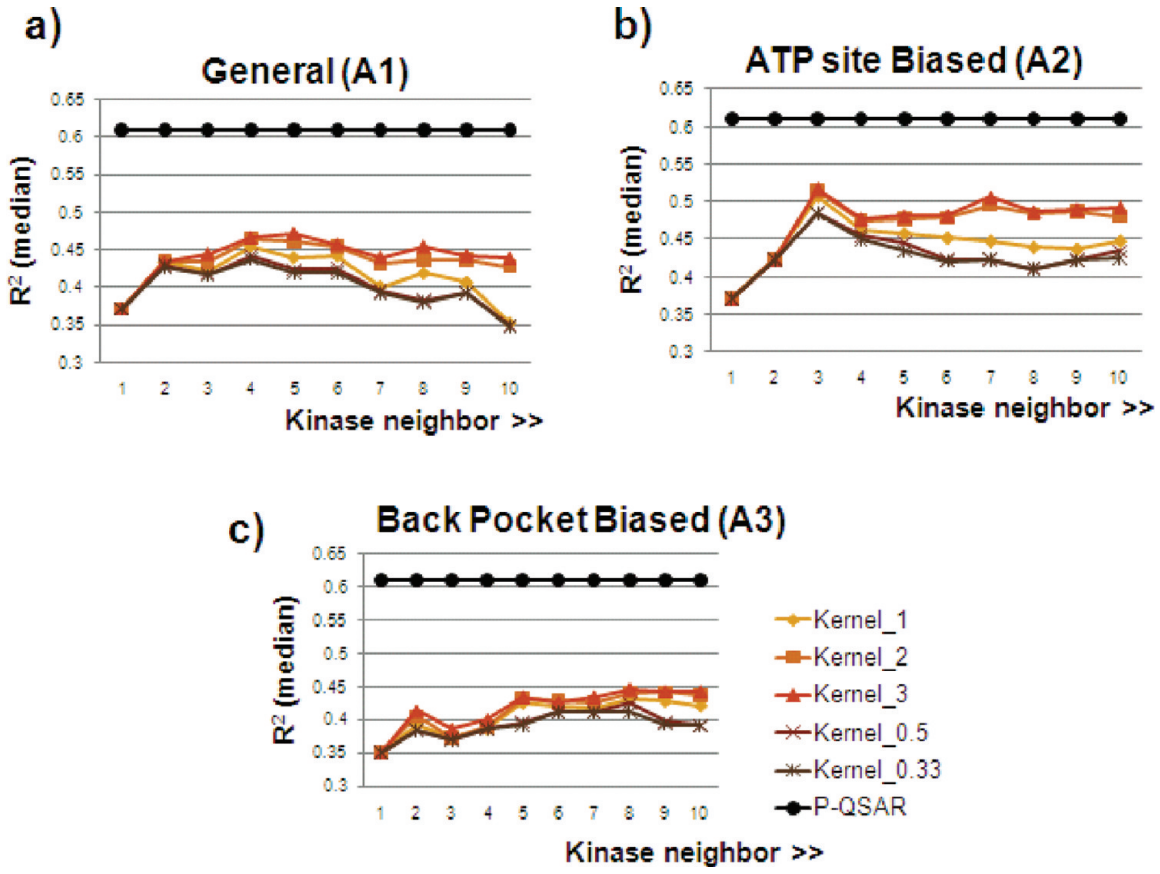


Figure 7. Prediction quality using three binding site definitions: (a) general, (b) ATP site biased, and (c) back-pocket biased. The *X*-axes plot the number of profile-QSAR neighbor models contributing to each prediction. The *Y*-axes are median R^2 over 51 kinase-kernel models for the best single combination of *b*, *m*, and *l* coefficients, from the combination of kinase neighbor count and binding site residue subset. Each plot has six series: the "gold standard" empirically parametrized profile-QSAR in black as reference and the series for the five distance-attenuation functions. In the legend, kernel_* refers to the polynomial used for the attenuation.

neighbors, understandable since distal neighbors have little weight in the average.

Figure 8 shows the response surface from the overall best-performing kinase-kernel interpolation, which used the ATP-site biased alignment, cubic attenuation function, and averages of three neighbors (Figure 7b). The response surface consists of 5151 points colored by the median R^2 for that particular combination of weights for binding site sequence identity (*b*), model quality (*m*) and ligand similarity (*l*); *b* is the most important single factor, but the best combinations also add a significant contribution of *l*. Similar proportions were consistent across all 135 combinations of parameters. These smooth mixture response surfaces are convincing evidence that the *b*, *m*, and *l* coefficients had been effectively optimized under these binding site definitions. However, ligand similarity proved unimportant after the binding site definition was fully optimized (see below).

Genetic Algorithm Optimization of Binding Site Residue Selection and Neighbor Count. Since binding site sequence identity was the most important coefficient, and the optimal neighbor count and performance depend on the binding site definition, the binding site definition and neighbor count were further optimized together. The number of possible residue combinations is enormous, so a GA[19] was employed to optimize the residue subset. The 46 residues in the union of all 3 binding site definitions were aligned for over 500 kinases to use both for model optimization and subsequent whole-kinome predictions. Each GA run used a cubic attenuation function and fixed number (*n*) of neighbors and only optimized the subset of residues in the binding site. The GA binding site definition or "chromosome" was a 46-bit string indicating the inclusion or exclusion of each of the 46 candidate residues. Thus, the GA varied both the identity and the number of residues in the binding site definition. The GA "fitness score" was the median R^2 for kinase-kernel predictions on the 51 kinases with existing experimental data and profile-QSAR models. The detailed GA workflow is illustrated in Figure 9. In step 1 the GA engine generates *n* binding site sequence subsets, or "individuals", for that GA "population", using the inputs from the previous iteration and applying mutation and crossover functions. In step 2, for each of the binding site sequence subsets (S_1−S_n) in the "population", the 51 × 51 binding site sequence identity matrix is generated. In step 3, for each binding site sequence subset (S_1−S_n), each of the 51 kinases are selected one at a time, and the *n* closest neighbors are identified for that kinase from the remaining 50, based on binding site sequence identity for that binding site sequence subset. In step 4, for each binding site sequence subset (S_1−S_n), for each of the 51 kinases, a kinase-kernel model is created using the neighbor information from step 3. In step 5, for each binding site sequence subset, each of the 51 kinase-kernel models are used to predict pIC_{50} for the set of compounds with experimental activity against that specific kinase.

dx.doi.org/10.1021/ci200314j |*J. Chem. Inf. Model.* 2012, 52, 156–170

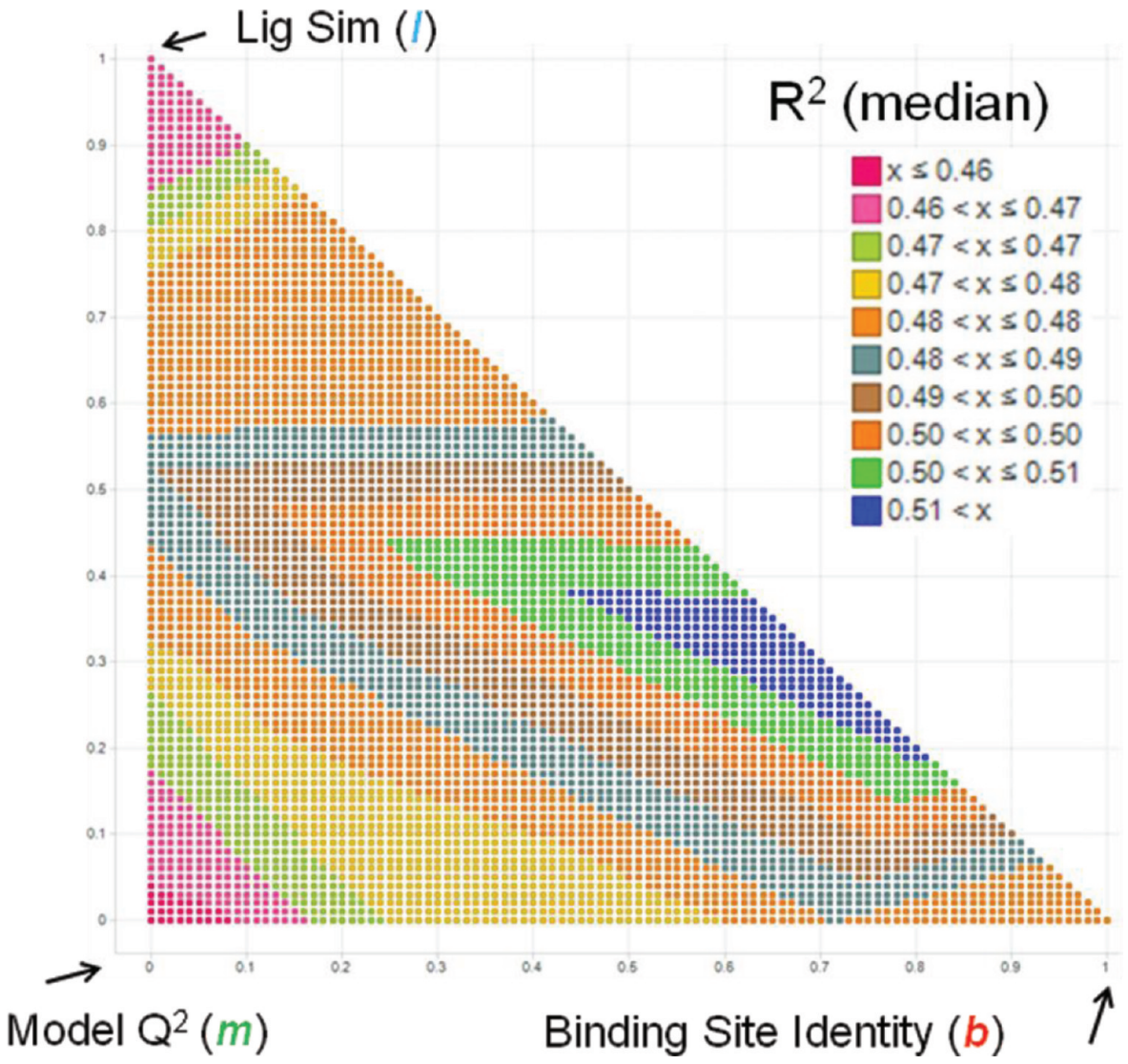


Figure 8. Mixture design response surface from the best performing kinase-kernel interpolation using the ATP site-biased alignment, cubic distance attenuation, and a neighbor count of 3.

In step 6, for each binding site sequence subset, a correlation squared (R^2) is calculated between the kinase-kernel pIC_{50} predictions and the experimental pIC_{50}s for each of the 51 kinases. In step 7, for each binding site sequence subset, the median R^2 is calculated using the 51 R^2 values from step 6. Each median R^2 serves as the "fitness score" for that binding site sequence subset or "individual". The "individuals" and their "fitness scores" are fed back to the GA engine to guide the mutation and crossover, generating a new population of binding site sequence subset definitions.

To collect a large sample of optimized kinase-kernel models, many runs were performed from many starting populations, varying the numbers of neighbors from 2 to 8. GA parameters were also varied to ensure good sampling. The first of two series of GA runs optimized the site residues for specified neighbor counts and GA population sizes. It consisted of 112 runs: 7 neighbor counts (2−8) by 4 population sizes (20, 40, 80, and 160) by 4 replicates. The plots in Figure 10 summarize some trends for the best kinase-kernel models from each of the 112 GA runs. Figure 10a,b is colored by the number of kinase neighbors used in the averages. Figure 10a plots the fitness score of the best model from each of the 112 GA runs on the *Y*-axis and the number of binding site residues (out of 46) in that model on the *X*-axis. Figure 10b is a histogram of fitness score colored by neighbor count. Two neighbor averages were consistently worst, and 6−8 were overall best. Two linear fit lines are drawn in Figure 10a: the black line uses fitness scores from all 112 GA runs, while the magenta uses only models from GA runs with neighbor counts from 3 to 8. While the former line does not show a slope, the latter has a negative slope, suggesting that parsimonious binding site definitions yield higher fitness scores. Figure 10c shows cumulative retrieval of the very highest fitness solutions, i.e., those with scores from 0.54 to 0.53, for the four population sizes. The population size of 160 retrieved fewer high fitness solutions than 20, 40, or 80. The scatter in these graphs also shows that these response surfaces are not smooth or simple.

The second set of GA runs focused on the best neighbor counts and population sizes determined from the first run, while varying mutation rate. It consisted of 108 runs: 3 neighbor counts (6−8) by 3 population sizes (20, 40, and 80) by 3 mutation rates (0.04, 0.06 and 0.08) by 4 replicates. The regression line in

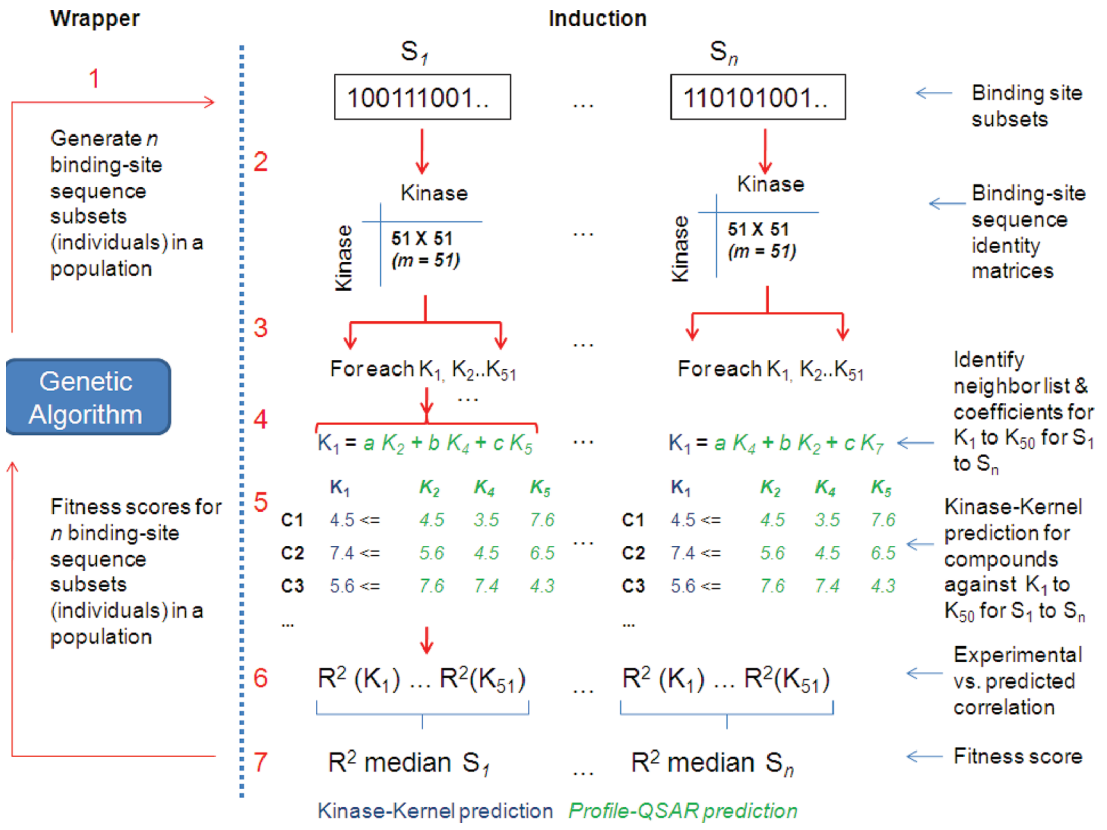


Figure 9. Workflow for carrying out the GA optimization of binding site sequence identity measure. On the left of the blue dotted line, the GA engine "wrapper" generates populations of binding site definitions for the "induction" phase, which evaluates the predictive "fitness score" of each generated binding site definition. In step 1, the GA applies mutation and crossover to the fittest members of the previous population of binding site definitions, to create a new population. For each binding site definition in the new population, a new binding site sequence identity matrix is generated (step 2). For each kinase in the matrix, *n* nearest neighbors are identified in step 3. In step 4, the nearest neighbor information is used to generate a kinase-kernel model for each unique combination of a kinase and binding site sequence subset. In step 5, the compounds with experimental pIC_{50} values against each of the 51 kinases are predicted using the kinase-kernel models specific to that kinase and binding site sequence subset. In step 6, for each binding site sequence subset, the correlation squared (R^2) is calculated using the experimental and predicted pIC_{50}s for each of the 51 kinases. The median R^2 or "fitness score" is calculated using the 51 R^2 values for each binding site sequence subset in step 7, and information is passed to the GA engine to guide the mutation and crossover, generating a new population of binding site definitions for the next generation.

Figure 11a recapitulates the trend that the optimized models had smaller subsets of residues in the site definition. The histogram in Figure 11b indicates a modest preference for the higher neighbor counts of 7 or 8. The histogram in Figure 11b indicates an advantage for the higher mutation rate of 0.08.

MOST RELEVANT RESIDUES

If the kinase-kernel models are capturing the physics of binding, rather than just curve fitting, then the residues most frequently selected by the GA for the best kinase-kernel models should be consistent with known kinase SAR and structural biology. The histogram in Figure 12a shows how many times each of the 46 residues occurs in the binding site definitions of the 234 best kinase-kernel models, i.e., those with fitness scores above 0.54, which all came from the second GA run. Interestingly, there is a sharp drop in frequency after the first 16 residues (highlighted by the cyan dotted box). These 16 "privileged" residues must carry the most information to predict SAR similarity and thereby identify the most relevant kinase neighbors for interpolating kinase activity. Figure 12b maps the 46 residues onto a PKA crystal structure (PDB 1QU8). The residues in red and orange occurred most frequently, while those in yellow, green, and cyan occur progressively less frequently. Blue residues were not among the 46. Note that this method does not pick up residues that are important but invariant between the 51 kinases, e.g., the DFG motif.

Figure 13 details the 16 privileged residues, mapped onto a kinase crystal structure of PKA (PDB 1QMZ). Magenta residue labels are based on PKA sequence numbers, with brief descriptions of their putative functional roles and PDB codes of archetypical kinase structures that exemplify these ligand interaction (in red) and protein stabilization (in blue) roles. In the hinge region, the GA identified the gatekeeper (GK) M120, the first hinge residue E121 (GK + 1), and the third hinge residue V123 (GK + 3). GK + 1 and GK + 3 form hydrogen bonds with the adenine group of the native substrate (ATP) and are the most important kinase pharmacophores for designing ATP competitive inhibitors. Gatekeepers[20,21] usually have hydrophobic side chains of varying bulk. Smaller gatekeepers allow access to the hydrophobic "selectivity pocket", which can be exploited to gain selectivity against off-target kinases lacking this kind of access due to bulkier side chains at this position.

Residues D127, L173, and T183, located in or around the ribose binding pocket, participate in ligand protein interactions,

dx.doi.org/10.1021/ci200314j |J. Chem. Inf. Model. 2012, 52, 156–170

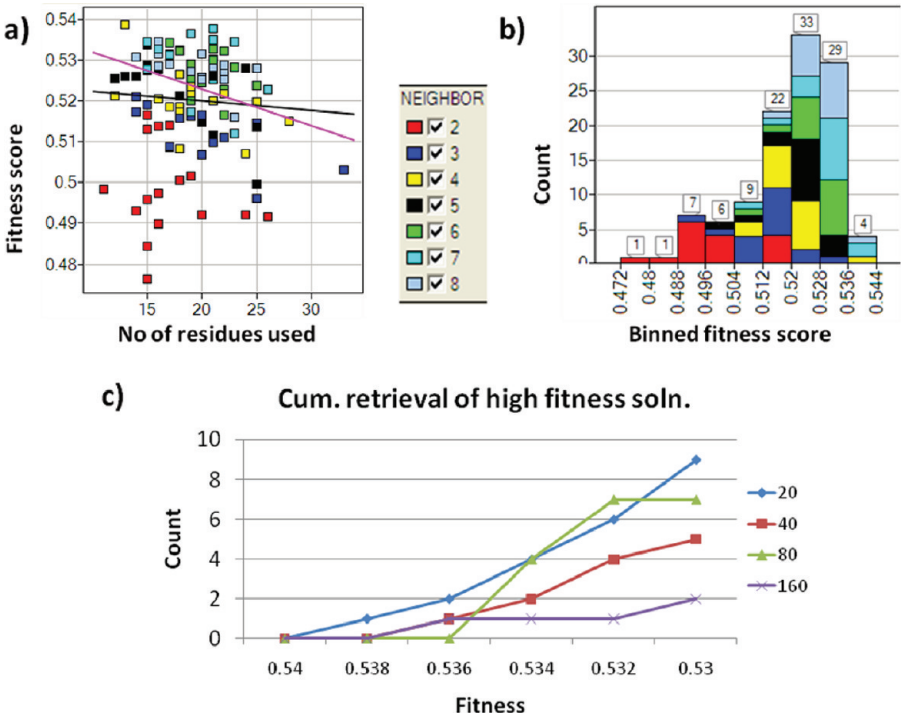


Figure 10. (a) A plot of the fittest models from 112 GA runs showing the fitness score on the *Y*-axis and number of residues in the sequence subset (out of 46) in the *X*-axis. (b) A binned histogram of the 112 fittest models colored by kinase neighbor count. (c) Cumulative histograms counting the number of models retrieved with scores above a given median R^2. Each series represents a different population size.

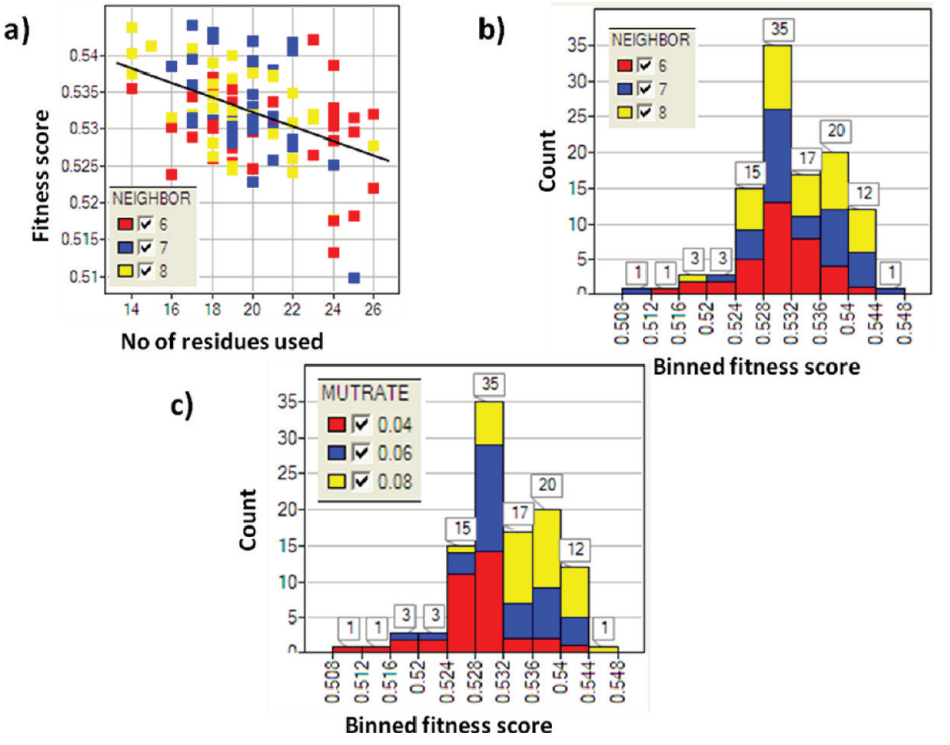


Figure 11. (a) A plot of fitness score vs number of residues in the binding site definition for the best kinase-kernel models from 108 GA runs. (b) A binned histogram of the 112 fittest models colored by kinase neighbor count. (c) A binned histogram of the 112 fittest models colored by mutation rate.

e.g., the residues corresponding to L173 and T183 interact with erlotinib in the costructure of EGFR in the "active" state (PDB code 1M17).[22] The carboxylate of D127 forms a charge−charge interaction with the cationic secondary amine

dx.doi.org/10.1021/ci200314j |*J. Chem. Inf. Model.* 2012, 52, 156–170

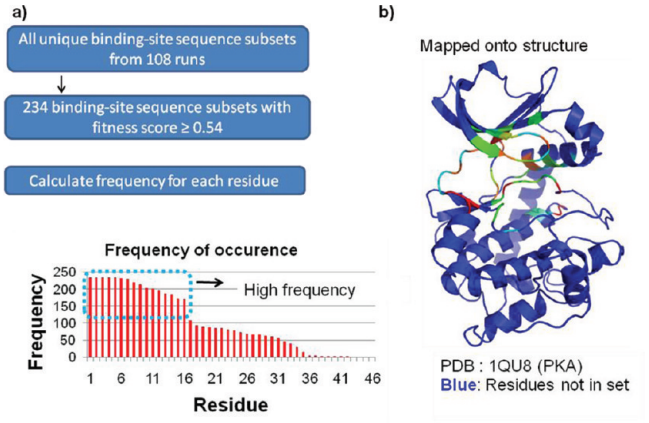


Figure 12. (a) Frequency of occurrence for each of the 46 residues in the 234 best kinase-kernel models (those with median R^2 for 51 kinases >0.54). (b) Mapping of the 46 residues onto a PKA crystal structure backbone ribbon. The 46 residues are colored by decreasing frequency of occurrence in the order red, orange, yellow, green, and cyan. Blue residues were not among the 46. The high-frequency residues are colored red or orange.

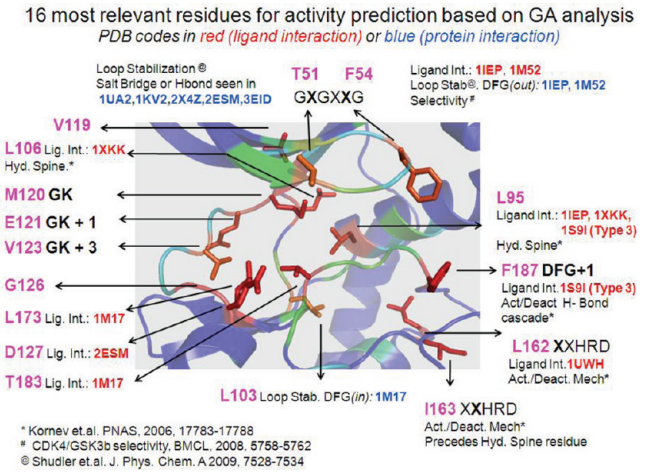


Figure 13. A close up view of the PKA binding site mapping the 16 most frequent residues (stick representation) in red or orange. Descriptions are given with each residue of their role in a ligand−protein interaction or protein stabilization as well as PDB codes of crystal structures where such a role is observed.

on the homopiperazine group of fasudil cocrystallized with ROCK-I kinase.[23] L103 lies at the floor of the binding pocket, and while it does not participate directly in a ligand protein interaction, it can aid in protein stabilization. For example, in the erlotinib−EGFR structure (PDB code 1M17), this residue stabilizes the "in" conformation through a hydrophobic interaction with the phenylalanine of the conserved DFG motif. Interestingly, Shudler et al.,[24] using their "BlockMaster" algorithm, have identified the region from 99 to 105 (PKA numbering), including L103, as being part of the "pivot block" which reaches to both the N- and C-terminal lobes and might play a functional role during the activation−deactivation process associated with changes in interlobe orientation.[25]

Two privileged residues are in the highly flexible G-loop region. Residue T51 is vectored out of the binding pocket and does not interact with the ligand. However, in a number of structures the corresponding residues are involved in a salt bridge or hydrogen bond with a residue in the same loop. These include CDK2 (PDB code 3EID), CDK9 (PDB code 1UA2), ROCK1 (PDB code 2ESM), P38 (PDB code 1KV2), and PAK (PDB code 2X4Z). Some other kinase structures, like EGFR (PDB code 1XKK, 1M17), cABL (PDB code 1IEP, 1M52), and bRAF (PDB code 1UWH), do not show these stabilizing interactions. The interactions rigidify the G-loop in these kinases. Several roles could be assigned to F54, the other privileged residue in the G-loop. The corresponding residue (Tyr) is involved in ligand−protein interaction in several c-ABL crystal structures (PDB code 1IEP, 1M52). The same residue also stabilizes the phenylalanine of the DFG motif in the "out" conformation through $\pi-\pi$ interactions in both structures. Shudler et al.[24] conclude similarly that the G-loop interacts with the residues of the DFG motif and appear to have lower B-factors in kinase structures in the "inactive" conformation than those in the "active" conformation. The flexibility of the G-loop is also likely to interplay with the "active/inactive" state of the enzyme. Stevens et al.[26] devised ligands selective for CDK4 over GSK3, by building hydrophobic contacts that project underneath the G-loop. They hypothesize based on available structures of CDK2 and GSK3b that the CDK4 residue corresponding to F54 is more conformationaly restrained and pointing "out", resulting in a pocket underneath the G-loop for the hydrophobic group to occupy. The same residue in GSK3b is more flexible and might be tucked "in" under the G-loop, thereby occluding the ligand.

Privileged residue L106 is located in the back pocket behind the gatekeeper. This residue interacts with the ligand in a cocrystal structure of EGFR (PDB code 1XKK). L106 along with privileged residue L95 (from the αC helix) form two of four residues that are part of the "hydrophobic spine",[27] which forms a centralized structural core that stabilizes the activated kinase conformation and is hypothesized to be an integral part of the activation−deactivation process. The residue corresponding to L95 also participates in ligand−protein interactions in c-ABL (PDB code 1IEP), EGFR (PDB code 1XKK) and the very unique MEK structure (PDB code 1S9I). 1S91 provided one of the first crystallographic evidence of an allosteric pocket that binds in a "type-3" mode, with no hydrogen bond in the hinge region. In this structure, the activation loop attains a unique conformation, pointing the privileged DFG + 1 residue, F187 into the pocket to form a ligand−protein interaction. In typical kinase structures in the "active" state, the DFG + 1 residue forms part of a β-turn with a hydrogen bond between the backbone of F (i) of the DFG motif and DFG + 2 (i + 3).

The DFG + 1 residue also forms a hydrogen bond through its backbone to the side chain of the Arg from the conserved HRD motif of the catalytic loop, which in turn is stabilized through an interaction with the phosphorylated activation loop residue. During the "active" to "inactive" transition, the cascade of changes in the hydrogen-bonding network includes a breaking of this β-turn, a rotation around the main chain connecting the DFG + 1 and DFG + 2 residues, breaking of hydrogen bonding between the DFG + 1 backbone and the Arg from the HRD motif, and a large backbone displacement of the DFG + 1 residue. Thus, the residue at this position likely affects the associated conformational changes, promoting a typical "inactive" conformation or perhaps even an atypical allosteric pocket.

Kornev et al.[27] identified H164 from the HRD motif as part of the "hydrophobic spine". The authors suggested that the adjacent privileged residue L163 might stabilize the R165 side chain

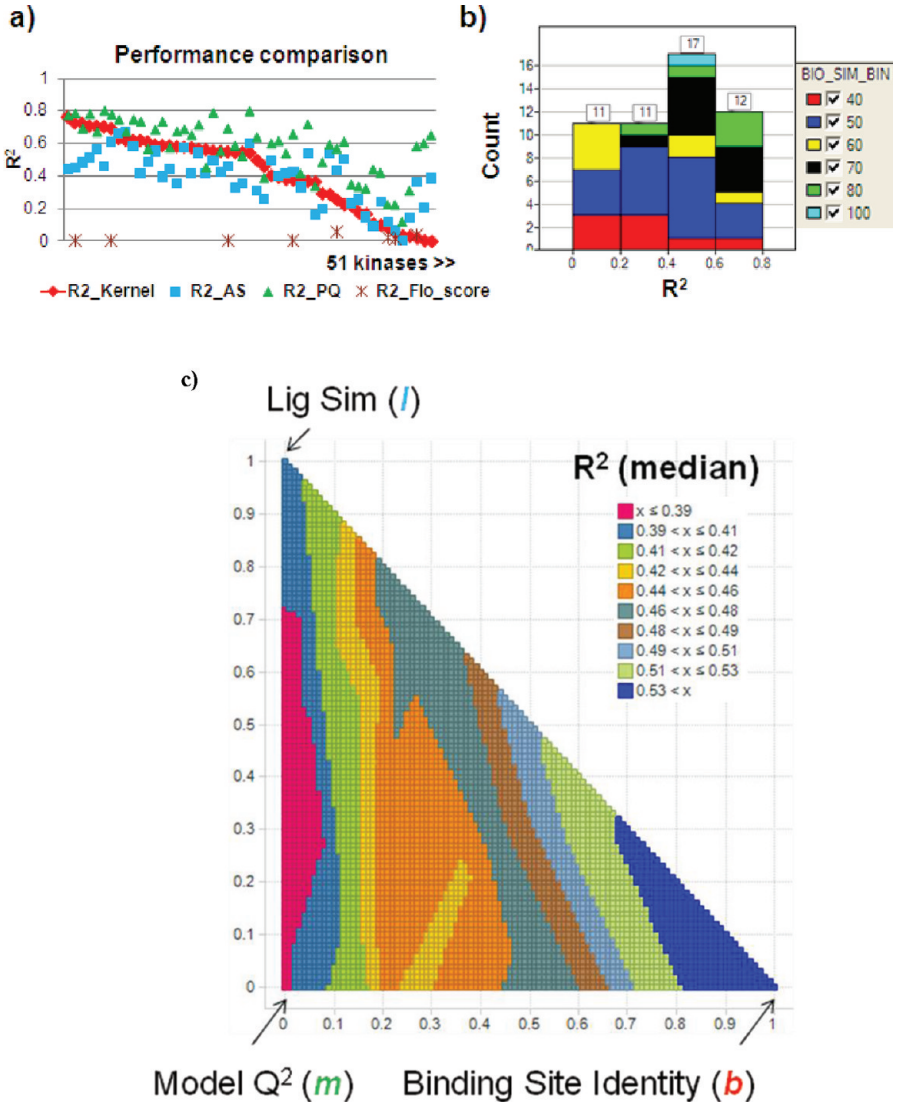


Figure 14. (a) A plot comparing performance for each of the 51 kinases by the global best kinase-kernel model (red), AutoShim (AS, cyan), profile-QSAR (PQ, green) and the Flo+ docking score where possible (brown). In this plot the R^2 is shown on the *Y*-axis and the 51 kinases, ordered in decreasing order of kinase-kernel performance, is shown in the *X*-axis. (b) A histogram of kinase-kernel models for the 51 individual kinases, binned into 6 groups by model R^2, colored by binding site sequence identity (BIO_SIM_BIN) of the second neighbor. (c) A mixture design response surface shows that kinase-kernel models using the global best site definition, neighbor count of 7, and cubic distance attenuation perform best without additional terms for model quality and ligand similarity.

through a hydrophobic interaction and play a role in coordinating the phosphorylation site geometry during the activation process. Privileged residue L162 participates in a ligand−protein interaction in bRAF structure 1UWH, as well as in type-3 inhibitors (internal data), and might also help stabilize the activation loop in the active state. No direct role could be assigned to privileged residues V119 adjacent to the privileged gatekeeper and G126 adjacent to privileged residue D127. However, specific pairwise correlations between privileged residues and their adjacent neighbors could well be important for activity prediction.

In summary, the 16 "privileged" residues that carry the most information for predicting SAR similarity include residues involved only in ligand protein interactions (M120, E121, V123, D127, L173, and T183) and residues hypothesized to play an important role in the loop dynamics and activation−deactivation mechanism of the kinases without interacting directly with the ligand (T51, L103, V119, G126, and I163) as well as some that might be involved in both (F54, L95, L106, F187, and L162). This suggests that SAR similarity requires similarity in conformational flexibility in addition to similarity in static interactions.

Performance Comparison. Figure 14a compares performance for each of the 51 kinases from the global best kinase-kernel model to the corresponding performance on the same assay data sets from surrogate AutoShim and profile-QSAR as well as with Flo+ docking scores for 9 kinases with crystal structures. The median R^2_{ext} for profile-QSAR is 0.61, while for the kinase-kernel models R^2 = 0.55. Compared to docking with a median R^2 of 0.01 or to any other method that does not require experimental training data, this is outstanding performance, approaching that of the profile-QSAR models from which the kinase-kernel predictions were derived. Surrogate AutoShim, an empirically parametrized target-tailored docking method, was

dx.doi.org/10.1021/ci200314j |*J. Chem. Inf. Model.* 2012, 52, 156–170

somewhat lower at 0.42. The histogram in Figure 14b for the 51 individual models, which is colored by binding site sequence identity of the second nearest neighboring kinase, shows that the best models tend to be for kinases with at least 2 highly similar neighbors (70−80% ID). This suggests that overall performance could be even further improved with additional experimental data to build profile-QSAR models for judiciously selected kinases.

The mixture design experiment (see above) was repeated using the global best residue subset from the GA optimization for the binding site sequence identity measure, along with model quality and ligand similarity. The cubic neighbor attenuation factor and a neighbor count of 7 were used. The triangular response surface (Figure 14c) shows that with the fully optimized binding site definition, pure binding site sequence identity alone is the best model weighing scheme.

External Validation. To demonstrate the prospective performance of the kinase-kernel models, 27 additional independent kinases were evaluated. Five of the 27 had sufficient data to also train profile-QSAR models, while the remainder did not meet the data set criteria for profile-QSAR model building, i.e., had fewer than 600 IC_{50}s and/or had fewer than 15 submicromolar IC_{50}s. For the 5 kinases with profile-QSAR models, kinase-kernel predictions were made on just the 25% held-out test sets for fair comparison with the profile-QSAR results. For the remaining 22 kinases, kinase-kernel predictions were made for all compounds. These prospective assay data sets ranged in size from 103 to 14 518 compounds, with dynamic ranges, measured as the standard deviation (σ pIC_{50}), varying from 0.27 to 1.45. These 27 data sets all originated from the same research laboratory as the original 51 assays and therefore shared some experimental similarities in assay conditions, constructs, etc. This consideration plays a role in the correlation between kinase-kernel predictions and experiment (see below).

As discussed earlier, the GA found 234 unique site definitions that gave fitness scores above 0.54. Collectively they corresponded well with a known kinase protein structure. These additional 27 prospective assays were enlisted both to evaluate performance on new kinases for which the kernel parameters had not been optimized and to ascertain how privileged the global optimum is among these 234 best binding site definitions and what GA fitness score constitutes an overall significant improvement. Tanimoto similarities were computed between the site definition bit string for the global optimum and the remaining 233 highest scoring site definitions. These were divided into 6 bins, spanning the range of similarities from T_c = 1 to 0.4. The best site-definition was selected from each bin. Kinase-kernel models were built for the 27 new kinases using the global-optimum site definition as well as these additional 5. R^2 dropped markedly when the dynamic range of the assay data, measured as σ pIC_{50}, fell below 0.8, so the assays were evaluated as 2 collections: the full set of 27 assay data sets (All_27) and the wide dynamic range subset of 18 assay data sets with σ $pIC_{50} \geq 0.8$ (18_GE_0.8). Figure 15a plots median R^2 for the global best and five other binding site definitions in order of decreasing similarity to the global optimum. Median R^2 for the wide dyanamic range assays ranged from 0.42 to 0.48, compared to 0.32 to 0.38 for the full set of 27 assays. Although not monotonically, prediction accuracy for both assay collections generally increased with similarity to the global optimum, even among these 234 most highly optimized binding site definitions, confirming a benefit from carefully optimizing the selection of binding site residues.

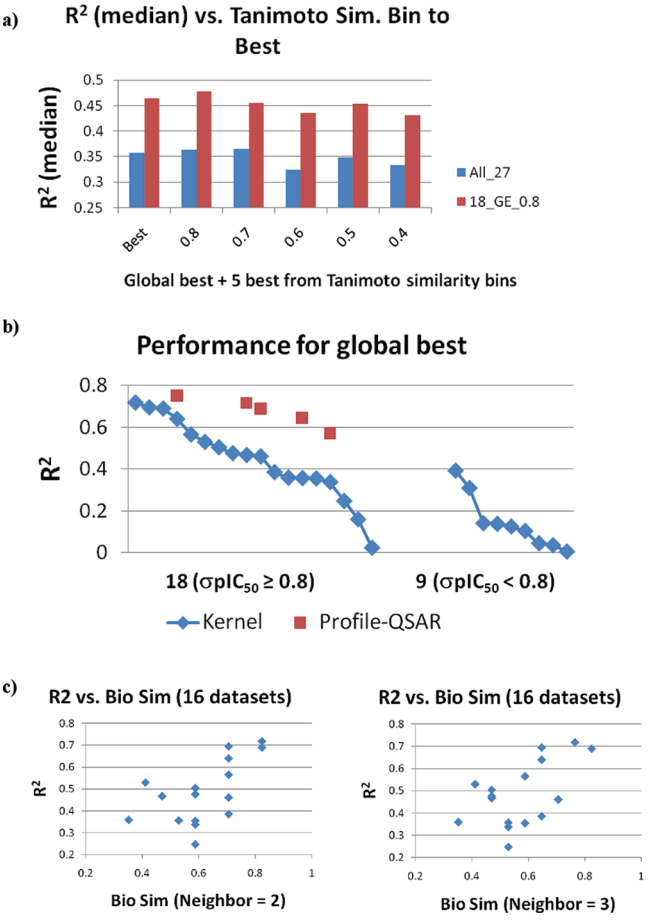


Figure 15. (a) A histogram showing the global best and the five highest fitness models from different Tanimoto similarity bins on the *X*-axis and median R^2 on the *Y*-axis. There are two series: All_27 in blue is the full set of 27 kinase assays, while 18_GE_0.8 in red refers to a subset of 18 kinase assays with a wider dynamic range. (b) A breakdown of the performance from the global best model. The 27 kinase assays are on the *X*-axis, divided into two groups of 18 with a wider dynamic range on the left and 9 with a smaller dynamic range on the right, and the R^2 is shown on the *Y*-axis. There are two series for the R^2 from kinase-kernel in blue and from profile-QSAR in red. (c) Scatter plots showing the kinase-kernel R^2 from the global best model on the *Y*-axis and the binding site sequence identity, "Bio Sim", of the second neighbor/third neighbor on the *X*-axis of the 16 higher performing assay data sets from the 18 with a wider dynamic range (see text).

Figure 15b shows kinase-kernel prediction performance for the 27 individual kinases using the global optimum binding site definition divided into two groups: the 18 with σ $pIC_{50} \geq 0.8$ and the 9 with σ $pIC_{50} < 0.8$. R^2_{ext} for profile-QSAR models is also plotted for the five cases with sufficient training data. Not surprisingly, the profile-QSAR models perform substantially better than the corresponding kinase-kernel models. This suggests a workflow of initial kinase-kernel virtual screening, followed by ordering and testing hundreds of predicted actives, and followed in turn by training a profile-QSAR model and a second round of virtual screening.

The two poorest performing kinase-kernel models from the better set of 18 share a common nearest neighbor kinase with a very low profile-QSAR Q^2 of 0.23. This neighbor is an isoform of the two target kinases, and compounds were specifically

dx.doi.org/10.1021/ci200314j |*J. Chem. Inf. Model.* 2012, 52, 156–170

optimized against it. This bias would be expected to confound the nearest neighbor SAR-similarity assumption. If these two outlier cases are removed, no correlation is observed between the R^2 from kinase-kernel prediction and the Q^2 from profile-QSAR prediction of the first-neighbor model for the remaining 16 assay data sets. This agrees with the conclusion from the mixture plot above that profile-QSAR model quality of the neighbors does not correlate with prediction accuracy. Figure 15c, plotting kinase-kernel model R^2 on the *Y*-axis vs binding site sequence identity of the second/third neighbor on *X*-axis, for these 16 assay data sets (minus the 2 outliers), shows a degree of correlation ($R^2 = 0.34$ for second neighbor and $R^2 = 0.31$ for third neighbor) between the two, reinforcing the conclusion that binding site sequence identity is the dominant factor.

An unavoidable confounding factor is the specific details of the kinase assays. Ideally, every well-optimized assay for a given kinase would produce very similar rank ordering of inhibitors. However, many factors can lead to lack of correlation between even well designed assays: specific protein construct (e.g., kinase domain, truncated domain, or full-length), and presence or concentration of cofactors, extent of post-translational modifications, such as phosphorylations, specific peptide substrate used, ATP concentration, etc. This underscores the hazards in assuming that biochemical assays recapitulate expected behavior in a biological context and raises uncomfortable questions about the relevance to drug discovery of optimizing against any biochemical assay or even any single cellular assay.[28] Since the kinase-kernel predictions depend on profile-QSAR models trained against specific assays for neighboring kinases, but not on any particular assay for the target kinase of interest, this also limits the success of any nontrained activity prediction method, including kinase-kernel models, for any given target kinase assay. The predictions might be excellent for one assay for a given kinase and poor for another. This bias was observed when 35 additional kinase assays were added from a different laboratory in a different geographical site in the company. Correlation between assays was poor for several kinases which had been assayed in both locations. By all objective measures, both assays were high quality and both had been used independently in active projects for chemistry optimization and antitarget profiling. The 35 assays from the remote location were large enough to construct high-quality profile-QSAR models. However, when these 35 new kinases were added to the original set of 51 as a source of neighbors for kinase-kernel modeling, their activity was only poorly predicted by kinase-kernel modeling, with median R^2 ranging from 0.19 to 0.22 for all 35 new assays from the remote site and 0.25 to 0.30 for 24 of those 35 assay data sets with σ $pIC_{50} \geq 0.8$. For profile-QSAR, median R^2 was 0.52 for all 35 assays and 0.54 for the subset of 24 assay data sets with σ $pIC_{50} \geq 0.8$. This profile-QSAR performance was lower than that for the original set of 51, with median $R^2_{ext} = 0.61$, although not poor enough to suggest a problem for the kinase-kernel modeling. This suggests that training models, like profile-QSAR, on kinase activity data for each individual assay tune the profile-QSAR model to the idiosyncrasies of that particular experimental procedure. Out of the 35 kinases assayed from the remote site, 17 also had assays from the local laboratory, 10 from the reference set of 51, plus 7 from the external test set of 27. For the local assays, median $R^2 = 0.31$, while for the remote assays, median $R^2 = 0.19$. This was despite the fact that the 17 additional assays were added to the 51 for the predictions from the remote lab. This suggests again that assay results are not just a function of the gene. Systematic differences between the assay protocols in different laboratories dictate that predicted results for future assays from a given laboratory are best interpolated from previous assays from that same laboratory. Kinase-kernel models only use a sequence-based measure of kinase similarity and implicitly assuming IC_{50} is a property of the sequence. With no training procedure to account for these systemic assay differences, like profile-QSAR, kinase-kernel predictions from one laboratory may not be transferable to assays from another laboratory in all cases.

Application Scenarios. The most straightforward use of kinase-kernel modeling is in projects where little or no starting data is available. In this case, the predictions from the default option, i.e., global best site definition with a neighbor count of 7, would be used in a virtual screen to select and prioritize an initial set of compounds for experimental screening. A second scenario arises when limited biological data are available for the target but are below the usual requirements for building a profile-QSAR model. A third scenario is to use kinase-kernel as an orthogonal method in addition to profile-QSAR and AutoShim in an iterative MTS. Since the kinase-kernel is not "trained" on data from specific chemotypes, it may sample chemical space differently from profile-QSAR and AutoShim. In the second and third scenarios, the availability of activity data allows for additional tuning of the kinase-kernel models. While the global best binding site definition with a neighbor count of 7 is the default choice, a different high-scoring binding site definition might give a better representation of the binding site similarities around specific target kinases. To allow the user some additional flexibility, the five diverse models discussed in the external validation studies above are available in addition to the global best, to evaluate which binding site definition provides the best correlation between experiment and kinase-kernel predictions.

Automated kinase-kernel predictions have been added to our automated kinase profile prediction workflow.[29] A panel of 92 profile-QSAR models is routinely applied to predict kinase activity profiles for the corporate archive and drug-like compounds from external vendor collections. The binding site sequence identity matrix for these 92 kinases generates predictions for the additional >400 human kinases for the corporate archive and vendor databases. If the profile-QSAR prediction matrix does not need to be updated, the kinase-kernel predictions for 4 million compounds across >400 kinases takes less than 30 min on the cluster.

Project Applications. So far, the kinase-kernel method has been applied prospectively to two kinase projects: "K19" and "K20". The K19 team could not screen the entire archive and wanted a set of rational orthogonal selection methods to select about 15 000 compounds toward a 50 000 compound MTS. About 11 000 compounds were selected from profile-QSAR and AutoShim models built from 1000 experimental IC_{50}s. Kinase-kernel was used as an orthogonal selection method to generate an additional prediction set. Predictive performance was evaluated using the global best and the five additional binding site definitions described above, with neighbor counts of 7 or 8. Figure 16a shows the predictive performance from the kinase-kernel variations. While the global best binding site definition with 7 neighbors performs well, with an $R^2 = 0.42$, the similar binding site definition (at $T_c = 0.8$), with 8 neighbors, performs better with $R^2 = 0.56$. Given that the comparison is based on 1000 unsupervised activity predictions, the difference cannot be explained by simple chance correlation. The later was therefore used to generate the selection of ~4000 additional compounds for the MTS.

K20 had a limited data set of 210 IC_{50}s. Although the profile-QSAR model had a reasonable $R^2_{ext} = 0.41$, the small data set size

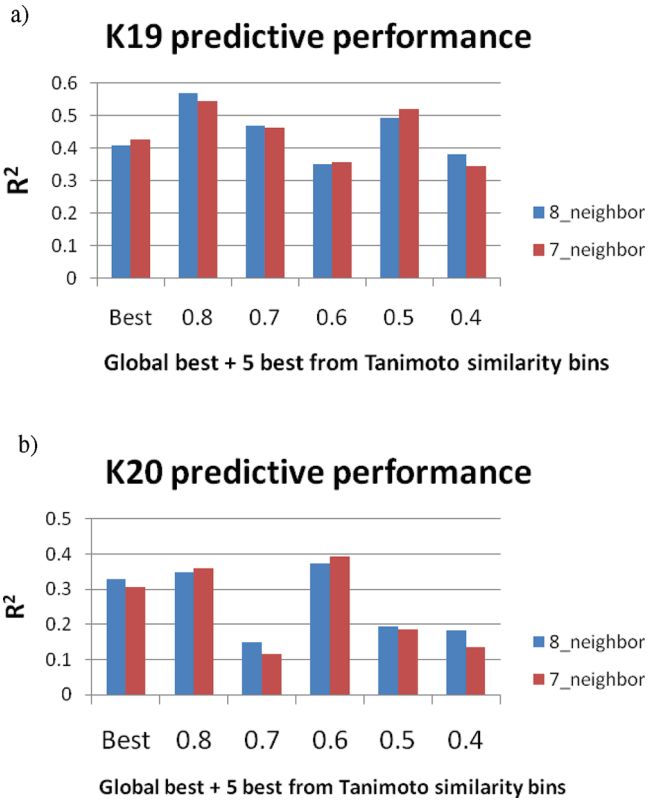


Figure 16. Predictive performance of kinase-kernel variations for kinases (a) K19 and (b) K20. The *X*-axis shows the global best and the five models selected from Tanimoto similarity bins, with R^2 on the *Y*-axis. The two series represent interpolation using seven or eight neighbor kinases.

and high proportion from just a few chemotypes caused concerns that selections were biased toward a limited area of the chemical space. Therefore, kinase-kernel was employed as an orthogonal selection method. A validation study (Figure 16b), similar to the one described for K19, used six binding site definitions and two neighbor counts of 7 or 8. The default option performed adequately with a $R^2 = 0.30$ but a neighbor with $T_c = 0.6$ with 7 neighbors performed better, with $R^2 = 0.39$. The later was used to generate a selection of 1000 compounds for further visual inspection and testing.

While in both cases, a solution that performed slightly better than the default could be found, the default itself performed at or above the lower predictive quality of $R^2 \geq 0.3$ that is often cited as productive for a screening application. While the percentage identity of target kinases with neighbors cannot be compared across different binding site definitions, it can be done for a given site definition. In this case, the percentage identities of the target kinase with the second neighbor, using the global best binding site definition, were 58% and 41%, respectively, for K19 and K20. These values are toward the lower end of the distribution from the 51 optimization kinases (Figure 14b) as well as the 16 test kinases (Figure 15c) and shows good prospective interpolation capability for difficult cases.

CONCLUSIONS

In silico alternatives to experimental HTS have many advantages: cost, speed, magnitude estimation, and coverage of commercial and virtual compounds. For the specific case of kinases, if training data are available, empirically parametrized profile-QSAR models can predict IC_{50}s with median $R^2_{ext} = 0.6$, accuracy approaching HTS IC_{50} experiments. Accurate profile-QSAR models have now been made for 92 of the 518 human kinases, covering most branches of the kinome. However, training these models requires hundreds of high-quality experimental IC_{50}s. Docking, the most common a priori virtual screening method, does not require any training data but does require a 3D protein structure, takes significant time, and the predicted affinity generally does not correlate with experimental IC_{50}s.[1] However, proteins from the same family, with homologous binding sites, often show similar SAR. The chemogenomic, kinase family based, "kinase-kernel" method described here predicts kinase activity without training data by taking a weighted average of profile-QSAR predictions from the previously modeled neighboring kinases with most similar ATP binding site sequences.

Three criteria were tested for weighting the neighbor contributions: (1) binding site sequence identity, (2) predictive quality of neighboring profile-QSAR models, and (3) similarity of the target compound with those in the neighbor's training sets. Only binding site sequence identity contributed when the binding site definition was carefully optimized. The best neighbor attenuation function fell off with the cube of similarity. A GA was employed to optimize the subset of kinase binding site residues that gave the best predictions. Several solutions with high but similar overall predictive power were identified. For the optimized kinase-kernel models, the median $R^2 = 0.55$, only modestly lower than median $R^2_{ext} = 0.61$ for the empirically parametrized profile-QSAR models on which the interpolations were based. A validation study was done on an additional set of 27 kinase assays not part of the initial set of 51 assays used for optimizing weighting factors. These kinase-kernel models performed well, with median $R^2 = 0.36$ for the full set of 27 assays, and particularly well for 18 assays with a wider dynamic range (σ $pIC_{50} \geq 0.8$) with median $R^2 = 0.48$. Analyzing the ensemble of GA solutions identified 16 privileged residues, from 46 candidates, which occurred in most of the best kinase-kernel models. These privileged residues, which correspond well with known kinase SAR and structural biology, provide reassuring validation that the model is physically realistic and not just curve fitting.

Combining the profile-QSAR and kinase-kernel models, activity was predicted for 2 million Novartis compounds plus an additional 2 million drug-like commercial compounds, for the >500 kinases covering the entire kinome. Thus, in silico screening for any profile of kinase activities has been reduced to a database lookup. The authors believe that models of this accuracy, covering an entire protein family with hundreds of pharmaceutically relevant targets, is unprecedented, and this database of 2 billion highly reliable predictions constitutes a unique and valuable resource.

While kinase-kernel predictions are very good, parametrized profile-QSAR predictions are significantly better. In addition, profile-QSAR was shown to reliably predict selectivity, with a median $R^2_{ext} = 0.6$ for nearly 1000 kinase pairs as well as entire kinase profile predictions for kinase-focused combinatorial libraries. Thus, an iterative approach is best employed. Based on initial kinase-kernel predictions, at least 400 compounds are selected, ordered, and their experimental IC_{50}s measured. Profile-QSAR, and AutoShim models are parametrized with these data, both for activity and selectivity prediction, and a second round of virtual screening is performed. Profile-QSAR models have already been applied to dozens of active Novartis therapeutic projects (manuscript in preparation). Two of these have now

included prospective kinase-kernel models. The results will be included in a forthcoming publication covering applications of iterative MTS.

■ ASSOCIATED CONTENT

Supporting Information. Two tables summarizing the targets, data sets, and statistical quality of the models. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: eric.martin@novartis.com.

■ ACKNOWLEDGMENT

P.M. would like to thank the NIBR education office for post doctoral funding.

■ REFERENCES

(1) 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.

(2) Martin, E. J.; Mukherjee, P.; Sullivan, D. C.; Jansen, H. Profile-QSAR: A Novel Meta-QSAR Method that Combines Activities Across the Kinase Family to Accurately Predict Affinity, Selectivity and Cellular Activity. *J. Chem. Inf. Model.* **2011**, *51*, 1942–1956.

(3) 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.

(4) Martin, E. J.; Sullivan, D. C. AutoShim: empirically correctedsS-coring functions for quantitative docking with a crystal structure and IC50 training data. *J. Chem. Inf. Model.* **2008**, *48*, 861–872.

(5) Vieth, M.; Erickson, J.; Wang, J.; Webster, Y.; Mader, M.; Higgs, R.; Watson, I. Kinase Inhibitor Data Modeling and de Novo Inhibitor Design with Fragment Approaches. *J. Med. Chem.* **2009**, *52*, 6456–6466.

(6) Vieth, M.; Sutherland, J. J.; Robertson, D. H.; Campbell, R. M. Kinomics: Characterizing the therapeutically validated kinase space. *Drug Discovery Today* **2005**, *10*, 839–846.

(7) Cheng, A. C.; Eksterowicz, J.; Geuns-Meyer, S.; Sun, Y. Analysis of Kinase Inhibitor Selectivity using a Thermodynamics-Based Partition Index. *J. Med. Chem.* **2010**, *53*, 4502–4510.

(8) Keiser, M. J.; Irwin, J. J.; Shoichet, B. K. The Chemical Basis of Pharmacology. *Biochemistry* **2010**, *49*, 10267–10276.

(9) Brylinski, M.; Skolnick, J. Cross-Reactivity Virtual Profiling of the Human Kinome by X-ReactKIN: A Chemical Systems Biology Approach. *Mol. Pharmaceutics* **2010**, *7*, 2324–2333.

(10) Kinnings, S. L.; Liu, N.; Buchmeier, N.; Tonge, P. J.; Xie, L.; Bourne, P. E. Drug discovery using chemical systems biology: repositioning the safe medicine Comtan to treat multi-drug and extensively drug resistant tuberculosis. *PLoS Comput. Biol.* **2009**, *5*, No.

(11) Tatonetti, N. P.; Liu, T.; Altman, R. B. Predicting drug side-effects by chemical systems biology. *Genome Biol* **2009**, *10*, 238.

(12) *Kernel methods*; http://en.wikipedia.org/wiki/Kernel_methods (accessed April 29, 2011).

(13) Cristianini, N.; Shawe-Taylor, J. *Kernel Methods for Pattern Analysis*; Cambridge University Press: Cambridge, U.K., 2004.

(14) Liu, W.; Principle, J.; Haykin, S. *Kernel Adaptive Filtering: A Comprehensive Introduction*. Wiley: Hoboken, NJ, 2010.

(15) *Nearest neighbor interpolation*; http://en.wikipedia.org/wiki/Nearest-neighbor_interpolation (accessed April 29, 2011).

(16) *Pyevolve genetic alogrithms*; http://pyevolve.sourceforge.net/ (accessed April 29, 2011).

(17) *NumPy*; http://numpy.scipy.org/ (accessed April 29, 2011).

(18) Mead, R. Response surface exploration. In *The Design of Experiments, Statistical principles for practical applications* **1988**.

(19) Goldberg, D. E. *Genetic Algorithms in Search, Optimization and Machine Learning*; Addison-Wesley Professional: Boston, MA, 1989.

(20) Schindler, T.; Sicheri, F.; Pico, A.; Gazit, A.; Levitzki, A.; Kuriyan, J. Crystal structure of HcK in complex with a Src family-selective tyrosine kinase inhibitor. *Mol. Cell* **1999**, *3*, 639–648.

(21) Wang, Z.; Canagarajah, B. J.; Boehm, J. C.; Kassisa, S.; Cobb, M. H.; Young, P. R.; Abdel-Meguid, S.; Adams, J. L.; Goldsmith, E. J. Structural basis of inhibitor selectivity in MAP kinases. *Structure (London)* **1998**, *6*, 1117–1128.

(22) Stamos, J.; Sliwkowski, M. X.; Eigenbrot, C. Structure of the epidermal growth factor receptor kinase domain alone and in complex with a 4-anilinoquinazoline inhibitor. *J. Biol. Chem.* **2002**, *277*, 46265–46272.

(23) Jacobs, M.; Hayakawa, K.; Swenson, L.; Bellon, S.; Fleming, M.; Taslimi, P.; Doran, J. The Structure of Dimeric ROCK I Reveals the Mechanism for Ligand Selectivity. *J. Biol. Chem.* **2006**, *281*, 260–268.

(24) Shudler, M.; Niv, M. Y. BlockMaster: Partitioning Protein Kinase Structures Using Normal-Mode Analysis. *J. Phys. Chem. A* **2009**, *113*, 7528–7534.

(25) Huse, M.; Kuriyan, J. The conformational plasticity of protein kinases. *Cell (Cambridge, MA, U. S.)* **2002**, *109*, 275–282.

(26) Stevens, K. L.; Reno, M. J.; Alberti, J. B.; Price, D. J.; Kane-Carson, L. S.; Knick, V. B.; Shewchuk, L. M.; Hassell, A. M.; Veal, J. M.; Davis, S. T.; Griffin, R. J.; Peel, M. R. Synthesis and evaluation of pyrazolo[1,5-b]pyridazines as selective cyclin dependent kinase inhibitors. *Bioorg. Med. Chem. Lett.* **2008**, *18*, 5758–5762.

(27) Kornev, A. P.; Haste, N. M.; Taylor, S. S.; Ten Eyck, L. F. Surface comparison of active and inactive protein kinases identifies a conserved activation mechanism. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 17783–17788.

(28) Shokat, K. M. Tyrosine kinases: modular signaling enzymes with tunable specificities. *Chem. Biol.* **1995**, *2*, 509–514.

(29) Martin, E. J.; Mukherjee, P.; Sullivan, D. C.; Kinnings, S. L. Profile-QSAR: A Novel Meta-QSAR Method that Combines Activities Across the Kinase Family to Accurately Predict Affinity, Selectivity and Cellular Activity. *J. Chem. Inf. Model.* **2011**, *51* (8), 1942–1956.