JOURNAL OF CHEMICAL INFORMATION AND MODELING

ARTICLE

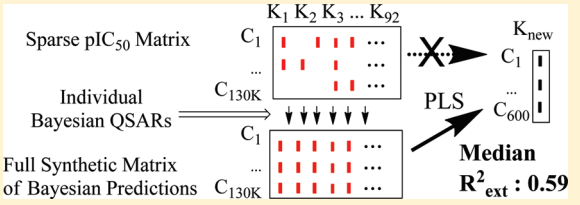
pubs.acs.org/jcim

Profile-QSAR: A Novel *meta*-QSAR Method that Combines Activities across the Kinase Family To Accurately Predict Affinity, Selectivity, and Cellular Activity

Eric Martin,* Prasenjit Mukherjee, David Sullivan,† and Johanna Jansen

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

ABSTRACT: Profile-QSAR is a novel 2D predictive model building method for kinases. This "*meta*-QSAR" method models the activity of each compound against a new kinase target as a linear combination of its predicted activities against a large panel of 92 previously studied kinases comprised from 115 assays. Profile-QSAR starts with a sparse incomplete kinase by compound (KxC) activity matrix, used to generate Bayesian QSAR models for the 92 "basis-set" kinases. These Bayesian QSARs generate a complete "synthetic" KxC activity matrix of predictions. These synthetic activities are used as "chemical descriptors" to train partial-least squares (PLS) models, from modest amounts of medium-throughput screening data, for predicting activity against new kinases. The Profile-QSAR predictions for the 92 kinases (115 assays) gave a median external R^2_{ext} = 0.59 on 25% held-out test sets. The method has proven accurate enough to predict pairwise kinase selectivities with a median correlation of R^2_{ext} = 0.61 for 958 kinase pairs with at least 600 common compounds. It has been further expanded by adding a "C_kXC" cellular activity matrix to the KxC matrix to predict cellular activity for 42 kinase driven cellular assays with median R^2_{ext} = 0.58 for 24 target modulation assays and R^2_{ext} = 0.41 for 18 cell proliferation assays.



The 2D Profile-QSAR, along with the 3D Surrogate AutoShim, are the foundations of an internally developed iterative medium-throughput screening (IMTS) methodology for virtual screening (VS) of compound archives as an alternative to experimental high-throughput screening (HTS). The method has been applied to 20 actual prospective kinase projects. Biological results have so far been obtained in eight of them. Q^2 values ranged from 0.3 to 0.7. Hit-rates at 10 uM for experimentally tested compounds varied from 25% to 80%, except in KS, which was a special case aimed specifically at finding "type II" binders, where none of the compounds were predicted to be active at 10 μM. These overall results are particularly striking as chemical novelty was an important criterion in selecting compounds for testing.

The method is completely automated. Predicted activities for nearly 4 million internal and commercial compounds across 115 kinase assays and 42 cellular assays are stored in the corporate database. Like computed physical properties, this predicted kinase activity profile can be computed and stored as each compound is registered.

INTRODUCTION

Experimental high throughput screening (HTS) is a key technology for identifying starting points for medicinal chemistry optimization.[1] However, HTS is extremely costly,[2] with a standard screen of a ~1.5 million compound archive costing nearly a million dollars and taking up to six months for completion.[3] Additionally, commercial collections and virtual libraries are not amenable to HTS, nor are certain assays, such as many sophisticated assays conducted on cell lines. Thus, alternative methods of hit discovery are wanted.

Virtual screening (VS), or in silico screening,[4−6] is one alternative to HTS. VS is not only faster and less expensive than HTS, but it can access chemistry outside the corporate archive. VS can be structure-based, wherein millions of "virtual compounds" are docked into an experimental protein structure or a model derived from structures of homologous proteins. It can also include ligand-based approaches such as QSAR models, pharmacophore models, and machine learning methods as well as similarity searching tools. With the advancement of computer hardware, these methodologies have attained the necessary throughput and have several success stories.[7] Some head-to-head comparisons[8,9] of VS and HTS on the same target of interest have shown that the methods can act complementarily, each recovering hits missed by the other. However, in routine evaluations, structure[10,11] and ligand-based methods[12,13] typically have only been shown to provide an enrichment of 2−7 times for the recovery of true positives. "Cherry-picking" even 10% of a large collection is prohibitively expensive. Furthermore, actual IC_{50} affinity predictions, which are important for prioritizing compounds for testing, rarely correlate well with experiment.

While general scoring functions in docking provide broad applicability toward a range of targets, greater accuracy can be achieved with empirically trained, target-specific scoring functions. In addition, instead of treating each new target as a unique idiosyncratic protein, a wider chemogenomic approach that

Received: December 20, 2010
Published: June 13, 2011

shares knowledge across target families can lead to tools of higher predictive power. These two assumptions underly the approaches to hit-finding by iterative medium-throughput screening (IMTS).

Iterative Medium-Throughput Screening. IMTS provides a high-throughput, accurate, and cost-effective marriage of virtual and experimental screening technologies, which harnesses the strengths of target-tailored scoring as well as chemogenomic relationships. These data-driven methods start with accurate medium-throughput IC_{50} data for a representative training set of molecules that sample the available chemical space. Target-tailored predictive models parametrized with these data are then used to virtually screen the entire corporate archive, as well as external databases, to generate a list of potential hits. Selected hits are assayed, and the resulting data may be appended to the original training set to carry the cycle iteratively forward, progressively exploring more uncharted but promising areas of the chemical space. While any QSAR method can be used for IMTS, two highly predictive IMTS methodologies specifically designed to take advantage of the structural relationships among members of a protein family have been developed: Surrogate AutoShim,[3,14] which has been previously described, and Profile-QSAR, a chemogenomic[2,15−17]/kinomics[18−27] method, which is the topic of this paper.

Profile-QSAR Compared to Previous Affinity Fingerprint Methods. Profile-QSAR is a 2D substructure-based method for building highly predictive models of new targets in a large protein family, in this case kinases, by incorporating an enormous database of related affinities from previously studied kinases. This "*meta*-QSAR" method models the activity of each compound against a new kinase target as a linear combination of its predicted activities against a large panel of 92 previously studied kinases comprising 115 assays. Profile-QSAR draws on ideas from several earlier technologies. The most obvious is "affinity fingerprinting",[28] in which each compound is characterized by its experimental affinity for a diverse panel of enzymes. Villar et al.'s pioneering TRAP methodology[29−31] characterized compounds in a large database by measuring experimental affinity for each compound against a panel of 18 functionally and structurally unrelated enzymes. They argued that the diverse binding sites acted as molecular calipers to measure the accessible shape and property profiles of the molecules. New query compounds could then be tested against the panel, and the database searched for compounds with similar fingerprints, which should therefore have similar shape and property profiles, and hence similar activity.

Profile-QSAR differs from the TRAP technology in 3 important ways: (1) TRAP intentionally uses a minimized set of functionally diverse enzymes, so a compound would not be expected to bind each enzyme in the same conformation or using the same pharmacophoric features. Profile-QSAR intentionally works within a family of enzymes that are expected to bind the ligands using essentially the same conformation and pharmacophores, or at most a small number of modes. Also, because it aims to maximize the amount of closely related experimental data from a protein family, it uses the largest available "basis set" of closely related enzymes, rather than a minimized basis set of diverse enzymes. (2) TRAP used experimental affinities directly, so new compounds of interest needed to be tested against the 18 enzyme reference set to get the descriptors needed for activity estimation. While Profile-QSAR models do require a large database of prior IC_{50} data to train the core Bayesian QSAR models, as well as hundreds of experimental IC_{50}s for the new target to train the PLS models, it requires no further experimental data to predict the activity of new compounds. This allows the screening of large compound databases or virtual libraries. (3) TRAP affinity fingerprints were used for database searching, rather than accurate prediction of biological activity. Bender et al. described Bayesian affinity fingerprints,[32] which use predictions from Bayesian models developed across a large panel of diverse enzymes to construct a virtual affinity fingerprint, obviating the need for experimental descriptors. Like TRAP, however, these descriptors were used for database searching rather than improving activity prediction.

Scores obtained from docking into a reference panel of protein targets have also been used in lieu of experimental activity to construct affinity fingerprints.[33] However, docking scores are generally very poor at predicting ligand affinities and rank ordering compounds.[34] The trends obtained through docking scores are sometimes better for a congeneric series of ligands, where the errors can cancel out, but become a serious limitation for diverse compounds in large chemical databases. Profile-QSAR was designed for VS of large internal and commercial databases and virtual libraries, for activity predictions against individual kinases or entire kinase profiles.

Thus, Profile-QSAR borrows from affinity fingerprint methods, the use of biological activity as chemical descriptors, but employs them to accurately predict IC_{50} for large or virtual compound databases within a homogeneous protein family.

PROFILE-QSAR THEORY

Profile-QSAR assumes that the properties of binding sites for a related family of proteins, such as the ATP binding sites of all kinases, can be approximately described as linear (or nonlinear) combinations of interactions with conserved features related in type, but differing in "magnitude", for each kinase: a hinge binder with 1, 2, or 3 hydrogen bonds, a larger or smaller gate-keeper, a deep or shallow hydrophobic pocket, a more or less flexible DFG loop, etc. In this case, the inhibition (pK_i or pIC_{50}) of a series of compounds for a new kinase (K_{new}) can be described by a linear (or nonlinear) combination of the affinities of a "spanning basis-set" of pK_i's from previously measured kinases as described in eq 1 and illustrated in panel (a) of Figure 1

$$pK_{i(x,new)} = \sum_{j=1}^{n} C_j \times pK_{i(x,j)} \quad (1)$$

where $pK_{i\,(x,\mathrm{new})}$ is the negative log of the inhibition constant for compound x against the new kinase. C_j is the coefficient for the jth basis-set kinase, and $pK_{i(x,\,j)}$ is the inhibition constant for compound x against kinase j. Internally, Novartis has a wealth of high-quality biochemical IC_{50} data against a range of kinases spanning all arms of the Sugen kinome[35] (Figure 1b). Envisioning this huge data set as a large "KxC" (Kinase X Compound) matrix (Figure 1c), there are nearly 130,000 compounds with measured activities for over 115 kinase assays. In principle, one could use experimental activities (K_1, K_2 etc) as compound descriptors to train a regression or machine-learning model that could predict activity of compounds against a new kinase of interest. However, there are two logistical problems: (1) Although the table includes 1.5 million IC_{50}s, it is still only 10% complete, and such a sparse matrix cannot be used for PLS model training. (2) It would not be advantageous to measure the activity of each new compound in up to 115 kinase assays in order to estimate activity for one new kinase. Adding an intermediate

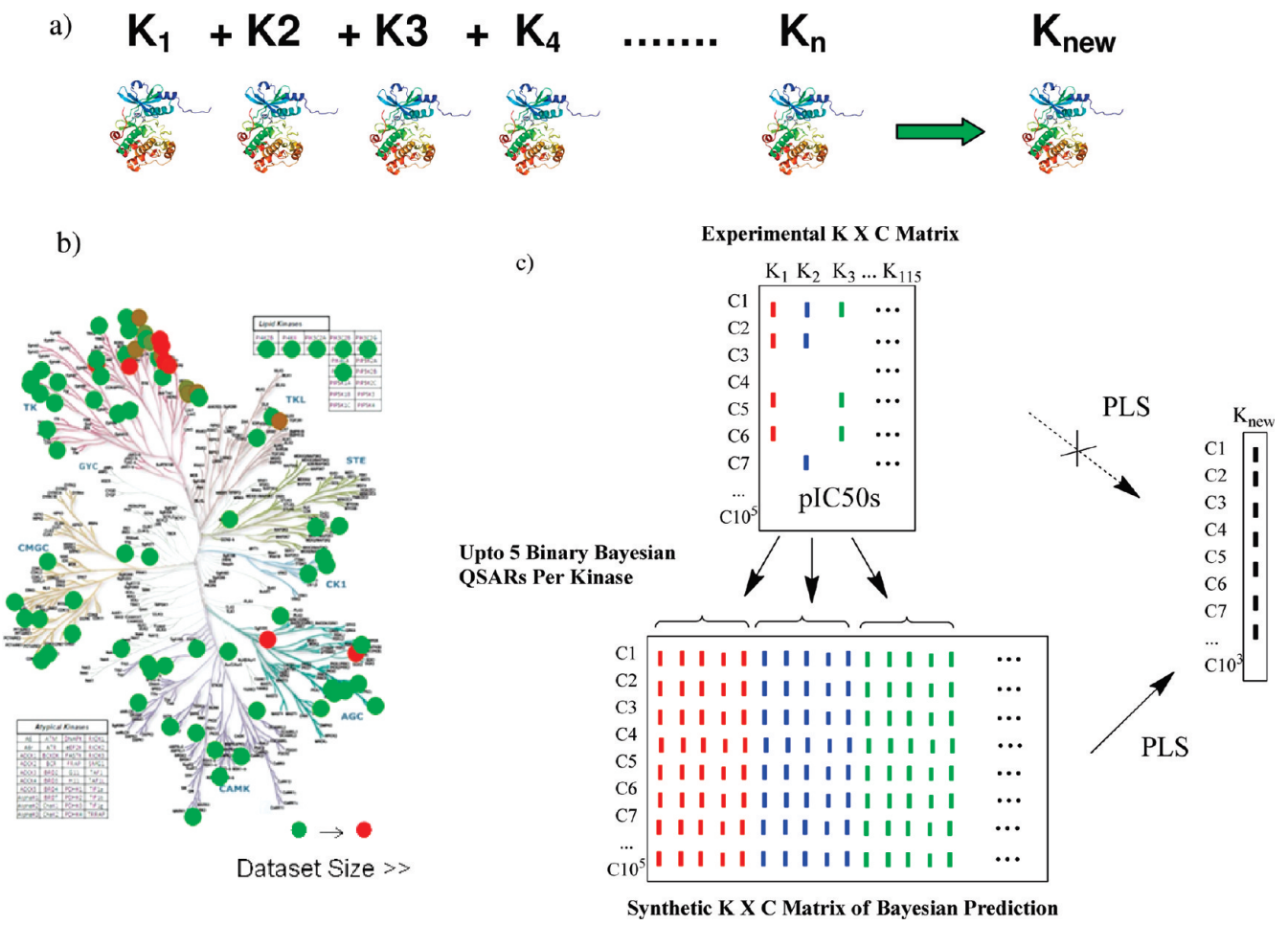


Figure 1. (a) Theory behind Profile-QSAR assumes that activity of a compound against a new kinase (K_{new}) can be modeled by a weighted sum of activities against a reference panel of previously studied Kinases (K_1 to K_n). (b) Distribution of 92 kinases across the arms of Sugen's human kinome dendrogram. The color scheme going from green to red indicates data sets of increasing size. (c) Schematic illustrating the Profile-QSAR methodology. At the top is a sparse experimental "KxC matrix" of kinase biochemical data, which is not directly suitable for PLS model building due to missing values. Therefore, first, going down the columns, up to five Bayesian QSAR models are trained on the available activity data for each individual kinase to generate a complete "synthetic" KxC matrix. The profiles of predicted kinase activities in this synthetic matrix are then used as "chemical descriptors" to construct a PLS model for K_{new}.

step of finding conventional QSARs for each individual basis-set kinase, as shown in panel (c) of Figure 1, solves both of these problems. First, going down columns of the matrix, the limited activity data available for each kinase are used to generate individual binary Bayesian QSAR models (see below) at up to five concentration thresholds for each assay. These models are used to generate a complete "synthetic" KxC matrix for all 130,000 compounds on all the 115 kinase assays. The predicted activities for the 115 assays in this synthetic KxC matrix are then used as the "chemical descriptors" to train a model against the IC_{50}s of the new kinase using PLS regression. Predictions on new compounds do not require any new experiments; each new compound is first run through the conventional Bayesian QSARs to fill in its row in the synthetic KxC matrix, and then run through the final PLS regression equation to predict its activity on K_{new}. All 1.5 million IC_{50}s contribute to each kinase model through this 2-step process, first combining the data down the columns in the Bayesian QSARs, then combining across the rows in the final PLS regression equation. These predictions prove to be much more accurate than the individual Bayesian QSAR models from which they were built.

The highly automated binary Bayesian QSAR method implemented in Pipeline Pilot (Accelrys, San Diego, CA) was selected to generate the synthetic KxC Matrix. Numerous modeling methods could be used for this purpose. Evaluations comparing continuous PLS QSAR models to binary Bayesian QSAR at up to five concentration thresholds showed comparable overall performance, but the former was far more computer intensive. The Naive Bayes[36] method uses Bayesian statistics to model binary data (active/inactive at a concentration threshold) as the dependent variable and a hashed bit-string of chemical substructures as inputs. It assigns conditional probabilities to the individual substructures from their frequency of occurrence in the active set compared to the whole data set or background. New molecules are assigned a probability of activity from a linear combination of the probabilities of its substructures.

METHODS

All workflows for the evaluation phase of this study were generated in Pipeline Pilot and could be run in an automated fashion. For higher throughput iterative screening applications,

Pipeline Pilot protocols launch shell and R[37] scripts on a Linux cluster for faster model building and predictions.

Data Preparation. The biological data for training and testing the models came from the Novartis proprietary kinase knowledge base, which can combine results from several assays to produce a single "best value" for each compound on each kinase. Each assay is given a "weight" reflecting its perceived quality. The "best value" is a weighted average pIC_{50} across all assays for that compound against that kinase. The weights are based on factors such as number of points in the curve, cofactor concentration, historical reliability of the screening laboratory, etc. Twenty-three kinases had two quite different assays with large data sets that either did not have substantial intersection to calculate a correlation or did not correlate very well. Lack of correlation among assays does not imply a problem. Assay results for a given kinase depend on many factors:[38] specific protein construct (e.g., kinase domain or full-length), presence or concentration of cofactors, extent of post-translational modifications such as phosphorylations, specific peptide substrate used, incubation time, ATP concentration, etc. This underscores the hazards in assuming that biochemical assays recapitulate expected behavior in a biological context. However, for extending the basis set for improved activity predictions, including the 23 additional assays that repeat some kinases under different conditions is not unlike including additional kinases to the basis set. It covers additional chemical space, improving the overall predictions across a diverse chemical archive. The additional columns were simply concatenated onto the KxC matrix for a total of 92 kinases but 115 assay columns. The cellular data were also extracted from the proprietary repository and then curated using automated Pipeline Pilot protocols to generate pEC_{50} values. Activities beyond the upper or lower limits of the assays range were offset by a factor of 10 in the appropriate direction to allow information from the inactive and extremely active compounds to be included.

Experience showed that the size and dynamic range (measured as standard deviation of pIC_{50}) of the training data sets correlated with the predictive quality of the models. Hence, models were only generated for kinases with over 600 data points and at least 15 compounds with $IC_{50} \leq 1\ \mu M$. Cellular assays are typically more demanding than biochemical assays, and the data set sizes were comparatively smaller. Therefore, a more lenient minimum requirement was imposed of over 400 EC_{50}s with at least 15 compounds having EC_{50}s $\leq 1\ \mu M$. This resulted in a total of 42 cellular activity data sets comprising 28,317 compounds with 52,130 EC_{50}s, a sparse matrix about 5% full. Kinase selectivity models were constructed between every pair of kinases with over 600 common compounds and at least 15 compounds with a $\Delta pIC_{50} \geq 3$ log orders. This resulted in 958 kinase pairs.

Bayesian QSAR. Model building commenced with the generation of the complete "synthetic activity" matrix using Pipeline Pilot's Bayesian QSAR.[39] As a compromise between binary and continuous modeling, up to five Bayesian categorization models were built for each kinase, using up to five concentration thresholds of 0.1%, 0.3%, 1%, 3%, and 10% of the pIC_{50} distribution for each kinase. Only the thresholds for which the actives set had 25 members or more were used. Thus, the 115 assays produced only 310 columns in the KxC matrix. The molecular descriptors used for the Bayesian QSAR were FCFP_6 fingerprints plus five additional physicochemical properties: aLogP, molecular weight (MW), number of hydrogen bond donors (HD), number of hydrogen bond acceptors (HA), and number of rotatable bonds (RB). To estimate the reliability of the Bayesian models, 75% of the data were used for training, and the remaining 25% were set aside as held-out test sets for a head-to-head comparison of the predictive power of the simple Bayesian QSAR models and the Profile-QSAR models.

PLS Model Building. The predicted Bayesian probabilities in the "synthetic activity matrix" for all the kinase assays were then used as chemical descriptors in PLS regression of K_{new} for the Profile *meta*-QSAR. Kernel-pls models were trained on the 75% training set using the pls package, in R. The number of latent variables (LV) was selected by 5-fold leave group out (LGO) cross-validation. "Q^2_{scaled}" was defined to penalize models with many LVs

$$Q^2_{scaled} = Q^2 - (0.002 \times LV)$$

Up to 25 LVs were allowed, albeit rarely needed, and the model with the highest Q^2_{scaled} was selected for prediction on the 25% held-out external test set. R^2_{ext} from the predictions on the test set was used for final model quality assessment.

Evaluation of Cross-Terms for Profile-QSAR. For each Profile-QSAR model, select quadratic and cubic terms were also added to the original "chemical descriptor" matrix of synthetic activities to test nonlinear relationships. Recursive partitioning (RP) on the original linear descriptor matrix was used to identify the most relevant cross-terms. For each set of training pIC_{50}s, six RP models were built using different thresholds, set at the top 1%, 5%, 10%, 30%, 50%, and 70% of the database. The rpart package in R was used, with a purity of 0.5 and tree depth of 3. Nonredundant quadratic and cubic cross-terms were collected and appended to the original descriptor matrix for PLS model generation.

Profile-QSAR on Cellular Activity Data. For modeling cellular activity, experimental pEC_{50}s for 42 cellular assays (C_kxC matrix) were concatenated onto the sparse experimental biochemical KxC matrix for a total of 157 experimental activity columns. Six computed physicochemical properties were also added to the resulting synthetic activity descriptor matrix for PLS regression to model cellular activity: cLogP (Daylight, Aliso Viejo, CA), MW, HA, HD, RB and topological polar surface area (PSA).

Profile-QSAR for Selectivity Predictions. Two approaches were compared to model Kinase selectivity: (1) "delta models", which were selectivity models trained directly on experimental ΔpIC_{50}s, which can only include those compounds with IC_{50}s on both the target and antitarget, and (2) "difference models", which use all the experimental data to build separate models for the target and antitarget, then subtract the two predicted pIC_{50}s. Delta models were built using 75% of the paired data and evaluated on a 25% held-out test set. For difference models, two individual Profile-QSAR models were trained using all the activity data except the 25% paired held-out data above. The pairs of individual predicted pIC_{50}s on the same paired 25% held-out test sets were then subtracted to calculate the difference model ΔpIC_{50}, which was compared against the experimental ΔpIC_{50} to evaluate the R^2_{ext}.

Automation. Two separate Pipeline Pilot protocols carry out the model building and prediction phases. The prediction protocol downloads structure/activity data from several Oracle databases. For qualified data beyond the upper or lower limits of the assay, a 10-fold shift is applied. Bayesian model generation is

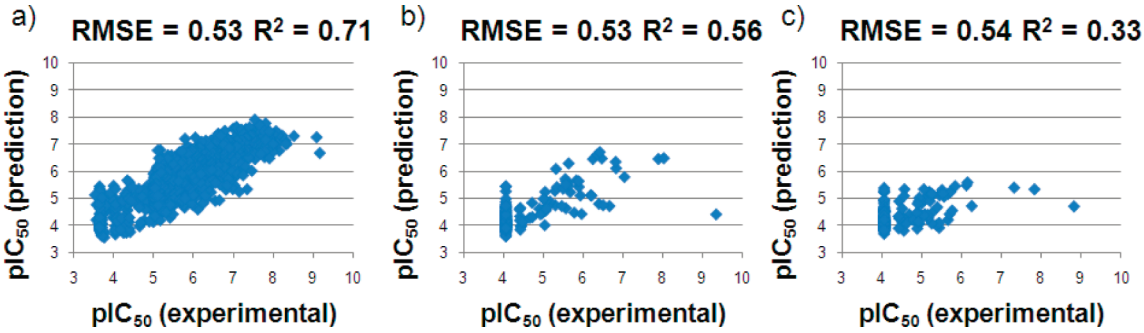


Figure 2. Scatter plots for the experimental and predicted pIC_{50}s from the test sets of three Profile-QSAR models. While all three have virtually identical RMSE, the R^2 decreasing from left to right parallels a clear decrease in ability to discriminate between active and inactive compounds. The test set sizes were (a) 1495, (b) 256, and (c) 249 pIC_{50}s. The percentage of the assay data sets that were at the highest concentration tested were (a) 2.3%, (b) 75%, and (c) 73%.

carried out as described above. The synthetic matrix is generated and transferred along with the activity data sets to a Linux cluster running standalone R. Pipeline pilot executes a high level shell script on the cluster that automatically creates individual job subdirectories, input files, and pbs job submission scripts that build the PLS models. An additional background script monitors for completed jobs, resubmitting lost jobs if required, and upon completion of model generation passes the results back to Pipeline pilot.

The prediction protocol starts by downloading chemical structures from the corporate archive and dividing them into batches of 100,000 compounds. The synthetic matrix is generated for each batch, which is similarly pushed to the Linux cluster. A shell script automatically generates and submits pbs scripts that predict activity for the entire corporate archive for each Profile-QSAR model. A background script checks for completion of jobs, resubmitting jobs as required. The individual predictions for all 115 biochemical and 42 cellular assays are joined into a full profile prediction table. Pipeline Pilot loads the predictions into the corporate database.

RESULTS

Performance of Profile-QSAR on Biochemical Data. Because of the distribution of the data, the chosen figure of merit was correlation, rather than a prediction residual, such as standard error. Most compounds in each assay are inactive at the highest tested concentration. To include these inactive data in model building, a 10-fold offset from the highest tested concentration is applied, i.e., if 70% of an assay data set have IC_{50} > 100 uM, these pIC_{50}s are all set to 3, while the remaining 30% might have measured pIC_{50} ranging from 4 to 9 (IC_{50} of 100 uM to 1nM). An inability to distinguish the few highly active compounds from the sea of inactive compounds might be disguised by correctly predicting the larger number of inactive compounds, while missing the compounds with high measured IC_{50}s. In the extreme case, a model that predicts that every compound has pIC_{50} = 3 would give a good standard error, but would do nothing to distinguish active from inactive compounds. On the other hand, correlation demonstrates a general agreement across the entire pIC_{50} distribution and therefore became the measure of choice. Figure 2 plots the predicted and experimental pIC_{50}s from the test sets of three Profile-QSAR kinase models. While all three have virtually identical RMSE, there is a clear loss in correlation going from left (R^2 = 0.71) to right (R^2 = 0.33). The test set sizes were (a) 1495, (b) 256, and (c) 249 pIC_{50}s. The percentage of the assay data sets that were at the highest concentration tested were (a) 2.3%, (b) 75%, and (c) 73%.

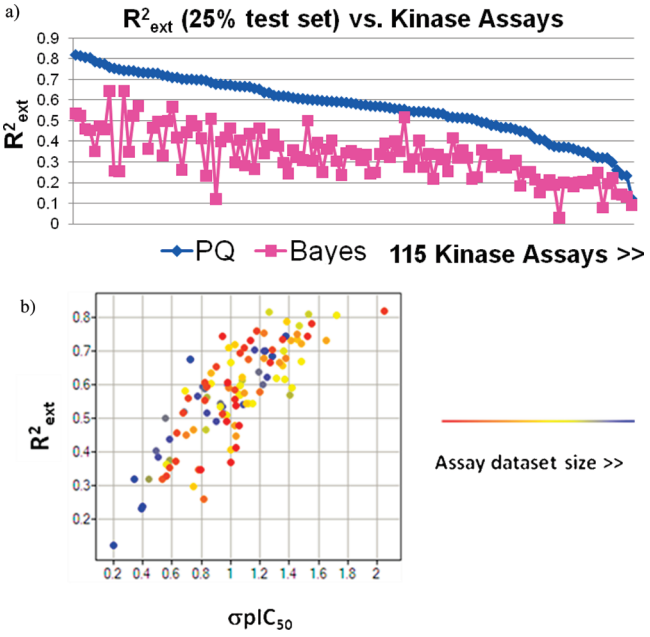


Figure 3. (a) Plot showing R^2_{ext} obtained by Profile-QSAR (blue diamonds) and simple Bayesian QSAR models (red squares) for the 25% held-out test set across 115 kinase assays. (b) Plot showing the correlation between R^2_{ext} from the 115 models with the dynamic range (σpIC_{50}) of the biochemical assay data. Colors indicate data set size.

A retrospective study evaluated what differences in R^2 values are significant. One hundred random 75/25 splits of two biochemical assay data sets were generated to evaluate how Q^2 and R^2_{ext} fluctuate with simple perturbations of the data set. One assay was from among the best performing models, and the other was near the median. For the high performing model, the median Q^2 for the 100 training set splits was 0.805, with standard deviation (SD) = 0.003, and R^2_{ext} = 0.804 with SD = 0.008 for the 25% held-out test set splits. For the more typical model, Q^2 = 0.612 with SD = 0.01, and R^2_{ext} = 0.616 with SD = 0.026. This suggests that differences of $R^2 \geq 0.02$ can be considered significant when comparing two models. However, while comparing median Q^2 or R^2_{ext} over many assays, i.e., for evaluating

dx.doi.org/10.1021/ci1005004 |J. Chem. Inf. Model. 2011, 51, 1942–1956

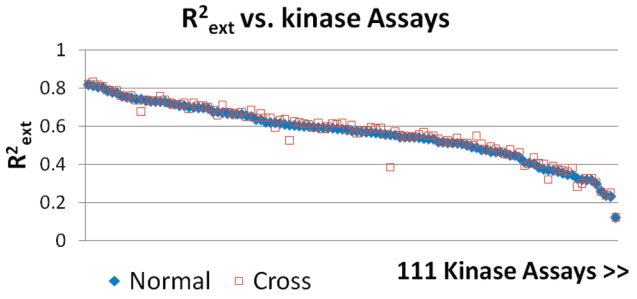


Figure 4. Plot showing the R^2_{ext} obtained by the "normal" Profile-QSAR (containing only linear terms, in magenta) and "cross" Profile-QSAR models (containing linear and cross-terms, in blue) for the 25% held-out test set across 111 kinase assays.

global performance across the kinome, the least significant difference will be much lower.

Profile-QSAR data sets ranged in size from 600 to about 58,000 IC_{50}s. The blue diamonds in panel (a) of Figure 3 show Profile-QSAR R^2_{ext} for 115 biochemical assays. The median R^2_{ext} was 0.59, indicating highly predictive models. The best few had $R^2_{ext} > 0.8$. A common rule of thumb implies models with $R^2 > 0.3$ are significant,[40,41] and 95% of the 115 Profile-QSAR models achieved this accuracy. Models with $R^2 \sim 0.3$ gave enrichments of about 40 fold, with typical retrievals of 25% of actives at 0.6% of the external test sets. The red squares show R^2_{ext} from the simple Bayesian QSAR models trained only on data for each specific kinase of interest. The median R^2_{ext} for these conventional QSARs is only a modest 0.32. In the majority of cases, Profile-QSAR yields a significant boost in predictive power over the Bayesian QSARs on which they are based, with a maximum gain of 0.7 units of R^2_{ext}, and only one case with no improvement. A plausible explanation for the improvement is that each Profile-QSAR prediction is informed by the entire experimental KxC matrix of nearly 1.5 million high-quality IC_{50} data points from around 130,000 chemical structures. The individual Bayesian QSARs combine all these data down the columns, and the final PLS combines the data across the rows. The quantitative relationships among correlated kinases that were measured on chemotypes which K_{new} had not seen, extends the model for K_{new} to indirectly include these substructures. One might anticipate that if one had as much experimental data for an individual kinase of interest, one might make QSAR models of even higher predictive power, but so much experimental training data is never available. As was shown previously for AutoShim, the quality of the generated models improves with greater dynamic range (measured as standard deviation of pIC_{50} values) and size of the training data sets, as shown in panel (b) of Figure 3.

Addition of Cross-Terms. One might anticipate that with such large and diverse experimental data sets, nonlinear models might perform even better. This was evaluated by adding selected quadratic and cubic cross-terms to the synthetic KxC matrix, generated based on RP of the synthetic KxC matrix at multiple thresholds (see Methods). Figure 4 plots the R^2_{ext} obtained for 111 models using PLS only on linear terms (normal) and those with additional cross-terms (cross). The median R^2_{ext} for the linear models was 0.58, while that from the models using cross-terms was 0.59. Because the addition of the cross-terms does not provide significantly better predictions, use of higher-order terms was abandoned for all subsequent applications. One possible explanation is that higher-order terms are already implicit in the synthetic KxC matrix. Principal components (PC) analysis on a Profile-QSAR prediction matrix of around 130,000 compounds and 115 kinase assays shows that 85% of the variance is explained by 26 of the 115 PCs. Thus, a minimal basis set could be created from a KxC matrix of only 26 carefully selected kinases. This does not suggest that a smaller kinase set should be used in Profile-QSAR. Rather it suggests that, in addition to improving the signal-to-noise, higher-order correlations among the additional redundant experimental data also contain enough implicit interactions to add the model flexibility that cross-terms might contribute.

Cellular Activity Modeling. A logical extension of Profile-QSAR trained on biochemical enzyme inhibition was to predict the cellular activity of kinase inhibitors. Cellular HTS assays are even more difficult than biochemical HTS assays, and accurate predictions are correspondingly more valuable. However, building accurate cellular QSARs is likewise more difficult because of the many additional factors influencing cellular activity: permeability,[42] differences in target and/or cofactor concentrations between biochemical and cellular conditions, different protein constructs, changes in protein conformation by virtue of complexation with protein partners, off-target effects, etc.[38] Among 42 cellular assays meeting the data requirements described in the Methods section, 24 were target modulation assays and 18 were cell proliferation assays. The target modulation end points were usually a specific phosphorylation of a downstream protein. The cell proliferation end points were quantifiable measures of cell viability. Target modulation assays are expected to be more specific for the target, or at least the pathway of interest, than the cell proliferation assays. Comparing panel (c) of Figure 1with panel (a) of Figure 5 highlights the modifications made to generate and evaluate the cellular models. A sparse activity matrix of cellular EC_{50} data (top right matrix in green in Figure 5a) is concatenated onto the sparse activity matrix of biochemical IC_{50} data (top left matrix in red in Figure 5a). Following the previous procedure, up to five Bayesian QSAR models were generated for each biochemical and cellular assay. Using the Bayesian QSAR predictions, a full synthetic matrix (bottom left matrix in Figure 5a) was generated comprised of both biochemical ("B", in red in the bottom left matrix from Figure 5a) and cellular ("C", in green in the bottom left matrix from Figure 5a) Bayesian predictions. In addition, six computed physicochemical parameters (see Methods) were added ("P", in blue in the bottom right matrix from Figure 5a) as additional descriptors for the PLS model building stage. The "B", "C", and "P" matrices shown in panel (a) of Figure 5 were joined in three combinations to generate different synthetic matrices of independent variables for Profile-QSAR evaluations: biochemical Bayesian + cellular Bayesian + physicochemical properties (B_C_P), biochemical Bayesian + cellular Bayesian (B_C), and biochemical Bayesian (B).

Panel (b) of Figure 5 shows the R^2_{ext} obtained on the 25% held-out test set from the 42 cellular assays. Target modulation assays are segregated from cell proliferation assays for comparison. Four different series are shown: B_C_P, B_C, B, and conventional Bayesian QSAR models. Given the difficulty of predicting cellular activity, the B_C_P series, which had the maximum number of descriptors, showed excellent performance with an overall median $R^2_{ext} = 0.52$ compared to the median $R^2_{ext} = 0.59$ obtained for the simpler biochemical activity models. All but three of the kinases gave R^2_{ext} greater than 0.3.

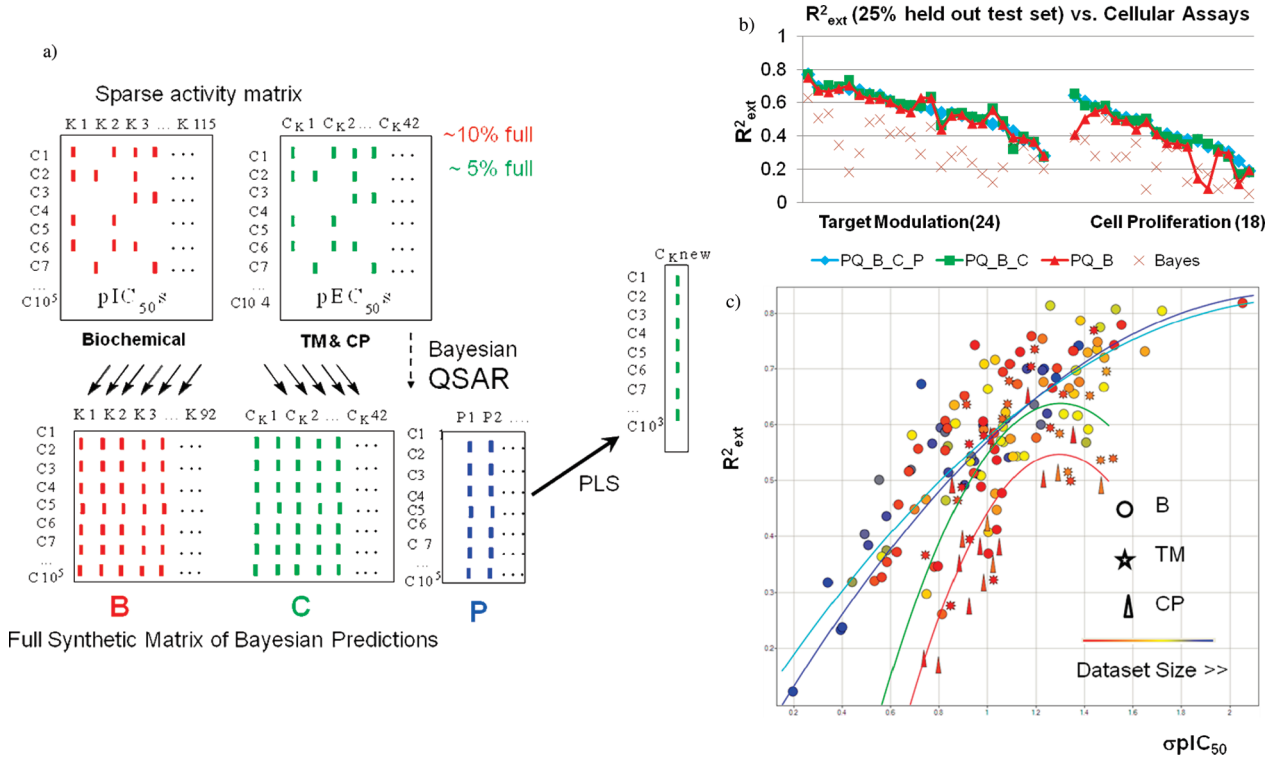


Figure 5. (a) Schematic of Profile-QSAR for modeling cellular activity of kinase inhibitors. A sparse activity matrix of 42 cellular pEC_{50}s in green is concatenated to the sparse experimental matrix of 115 biochemical kinase pIC_{50}s in red. As in Figure 1, each biochemical or cellular assay trains Bayesian QSAR models at up to five concentration thresholds, which in turn generate a synthetic activity matrix of Bayesian predictions. The portion of the Bayesian synthetic activity matrix coming from the biochemical Bayesian QSAR models is called "B" (in red), while the portion coming from cellular Bayesian QSAR models is designated as "C" (in green). Six additional computed physiochemical properties are shown in blue and referred as "P". The B, C, and P descriptor blocks were evaluated in various combinations as chemical descriptors to build PLS models of new cellular activity assays. (b) Plot showing R^2_{ext} (25% held-out test set) on the *Y*-axis for 42 cellular assays on the *X*-axis. The assay data sets have been segregated into target modulation on the left and cell proliferation on the right, showing that target modulation assays yield better models. (c) A plot showing R^2_{ext} on the *Y*-axis vs dynamic range, measured as $\sigma(pIC_{50})$, plotted on the *X*-axis. Circles are biochemical models, stars are target modulation cellular models, and triangles are cell proliferation models. Color, from red to blue, shows increasing data set size. The four lines are quadratic fits to subsets of the models: cyan — all biochemical data sets, blue — biochemical data sets smaller than the largest cellular data set, green — all target modulation data sets, red — all cell proliferation data sets.

Surprisingly, the performance did not drop for the B_C series, which lacked the calculated physicochemical properties, with an identical median R^2_{ext} of 0.51. This does not imply that the physicochemical properties are unimportant but rather that they are implicitly encoded by linear combinations among the biochemical and cellular activities in the activity descriptor matrix. Even the overall performance of the B series containing only biochemical activity was only slightly lower than B_C_P and B_C, with a median R^2_{ext} of 0.49. Models with the additional cellular Bayesian descriptors (B_C) outperform the corresponding B-only models mainly for six particular cell proliferation assays, with an improvement of 0.05 R^2_{ext} units or greater. Presumably, many factors that also influence cellular activity, such as permeability, subcellular localization, types and concentrations of cofactors, and perturbations to the binding site itself, are already encoded in differences among the biochemical assays. In addition, the current KxC matrix has only 20% as many cellular EC_{50}s as biochemical IC_{50}s, limiting their influence on the models. The fourth series shows the performance of the simple Bayesian QSAR models, which had a median R^2_{ext} of only 0.29, again illustrating Profile-QSAR's significant boosts in predictive power.

Panel (b) of Figure 5 shows that the 24 target modulation assays with median $R^2_{ext} = 0.58$ gave much better models than the 18 cell proliferation assays with median $R^2_{ext} = 0.41$, supporting the hypothesis that Profile-QSAR works by capturing the range of kinase activity possibilities, not just curve-fitting. While adding cellular data did not always improve predictions, using the B_C matrix for cellular activity modeling was chosen as the default for cellular activity modeling. The PLS technique generally performs well with multicorrelated variables. Hence, adding cellular assay columns did not diminish performance. A strict criterion was used in adding latent variables (see Methods) to prevent overfitting. While historically cellular assays have been conducted at a much smaller scale and as a follow up to biochemical screening, high-throughput cellular assays are increasingly used, so the future potential for a more significant contribution from cellular data should increase.

Panel (c) of Figure 5 plots R^2_{ext} of the biochemical, cell proliferation, and target modulation assays against the dynamic range (σpIC_{50}) of the training sets, which was shown in panel (b) of Figure 3 to have a large influence on biochemical R^2_{ext}. The cyan, green, and red quadratic fit lines to the biochemical, target modulation, and cell proliferation assays, respectively, show a progressively poorer correlation. Biochemical data sets tend to be larger, and larger training sets also tend to produce better models, as shown by the blue circles which tend to lie above the cyan fit

dx.doi.org/10.1021/ci1005004 |*J. Chem. Inf. Model.* 2011, 51, 1942–1956

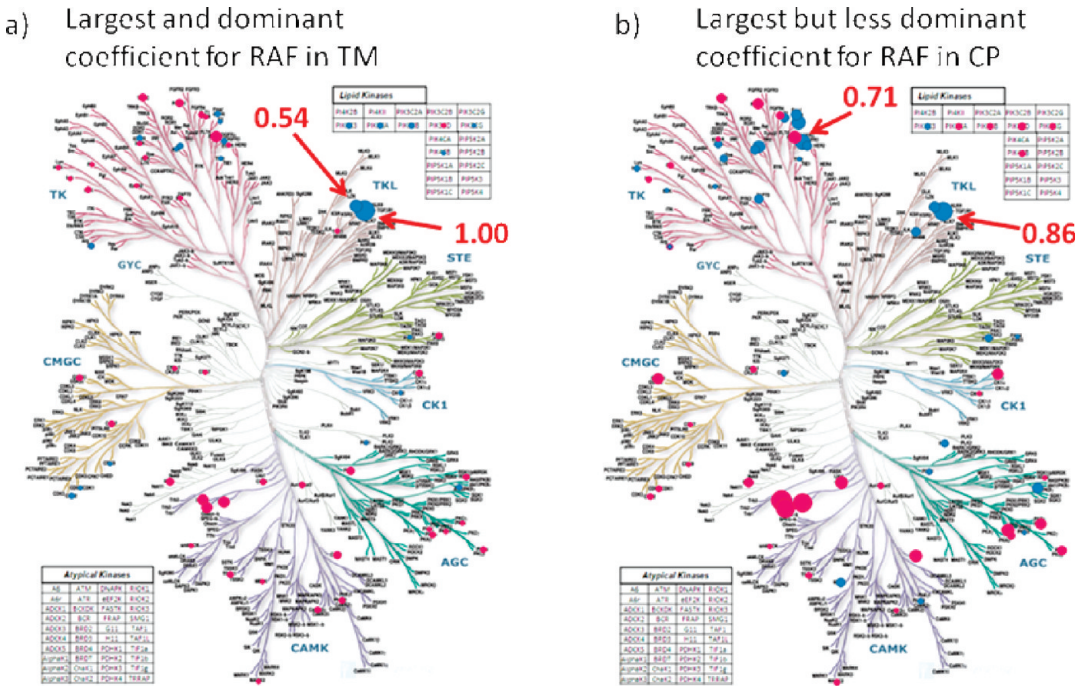


Figure 6. Kinome trees showing kinase biochemical IC_{50} descriptors used for studying the contributions of biochemical kinase assays to the (a) target modulation and (b) cell proliferation cellular RAF assays. The size represents the magnitude of the coefficient, blue being positive and red being negative.

line for biochemical activity models. However, a second quadratic fit (blue line) using only biochemical activity models with training sets smaller than the largest cellular training set still lies above those for cellular assays, indicating that the smaller cellular training set sizes alone do not account for the lower accuracy.

Cellular Profile-QSAR Coefficient Analysis. Although cellular Profile-QSAR was developed primarily for IMTS, the model coefficients can also suggest which kinases might be responsible for cellular activity. Figure 6 plots the kinase coefficients of Profile-QSAR models, plotted on the Sugen kinome tree, for a RAF target modulation and a "RAF-driven" cell proliferation assay that did not correlate with each other. The experimental biochemical and target modulation assays correlated better (R^2 = 0.44, 2303 compounds) than the experimental biochemical and cell proliferation assays conducted on the same cell line (R^2 = 0.24, 1657 compounds). Even though the cell line was supposedly "driven by" RAF kinase, inhibition of other kinases might also impact cell proliferation. These PLS models used only the biochemical assay block to simplify interpretation. The diameter of the circle corresponds to the coefficient magnitude for that kinase, while color indicates the sign (blue = +; red = −). For the target modulation assay, the coefficient for bRAF is almost twice as large as any other kinase, and cRAF is second, suggesting that the target modulation end point was driven largely by RAF inhibition. The cell proliferation plot shows that while bRAF comes up as the largest single positive coefficient (relative value of 0.86), PDGFRb, also has a large positive coefficient (relative value 0.71), as do other Tyrosine kinases (TK). These kinases might also be responsible for inhibiting cell proliferation. To determine the magnitude of meaningful coefficients, 200 cell proliferation and target modulation models were made from 100 random 50/50 splits of the assay data. For each 50/50 split, the ratio between the mean (absolute value) and the standard deviation for each descriptor was computed. For the RAF target modulation assay, bRAF and cRAF IC_{50} had the two most stable coefficients with stability scores of 26.4 and 23.48, respectively. For the cell proliferation assay they were the 34th and 40th most stable, with values of 6.76 and 5.4, respectively. Several TKs ranked higher in stability in the case of cell proliferation, e.g., PDGFRb was ranked 10th with a coefficient stability score of 12.62. The possible role of PDGFRb is supported by a recent report of a selective dual RAF/PDGFRb inhibitor.[43] Additionally, PDGFRb up regulation has also been recently identified as a RAF resistance mechanism in the clinic[44] and combination therapy[45] with RAF/PDGFRb dual inhibitor has been suggested as possible clinical solution to resistance. Thus, Profile-QSAR analysis predicted a multikinase dependency that was later validated through clinical observations.

Selectivity Modeling. Selectivity modeling was tested using 958 kinase assay pairs with at least 600 tested compounds in common. Panel (a) of Figure 7 shows the predictive performance of the models on 25% held-out external test sets. As mentioned in the Methods section, selectivity was predicted both by subtracting predicted pIC_{50}s from the individual biochemical activity Profile-QSAR models (difference models) and by training 958 new Profile-QSAR models directly using experimental ΔpIC_{50} as the dependent variable (delta models). Difference models might give better results where the models can be trained on more compounds, especially if few common compounds were tested against both kinases in the pair. The latter might have less noise with ample paired data. As explained, all kinase pairs in this test had substantial paired data to be able to accurately evaluate the performance. Overall, Profile-QSAR delta models built directly on experimental ΔpIC_{50} performed slightly better, with a median R^2_{ext} = 0.61 versus a median R^2_{ext} = 0.58 for difference models. The correlation for corresponding Bayesian models trained directly on experimental ΔpIC_{50} was much lower, at a median R^2_{ext} = 0.36, showing the benefit of Profile-QSAR. Among the

dx.doi.org/10.1021/ci1005004 |J. Chem. Inf. Model. 2011, 51, 1942–1956

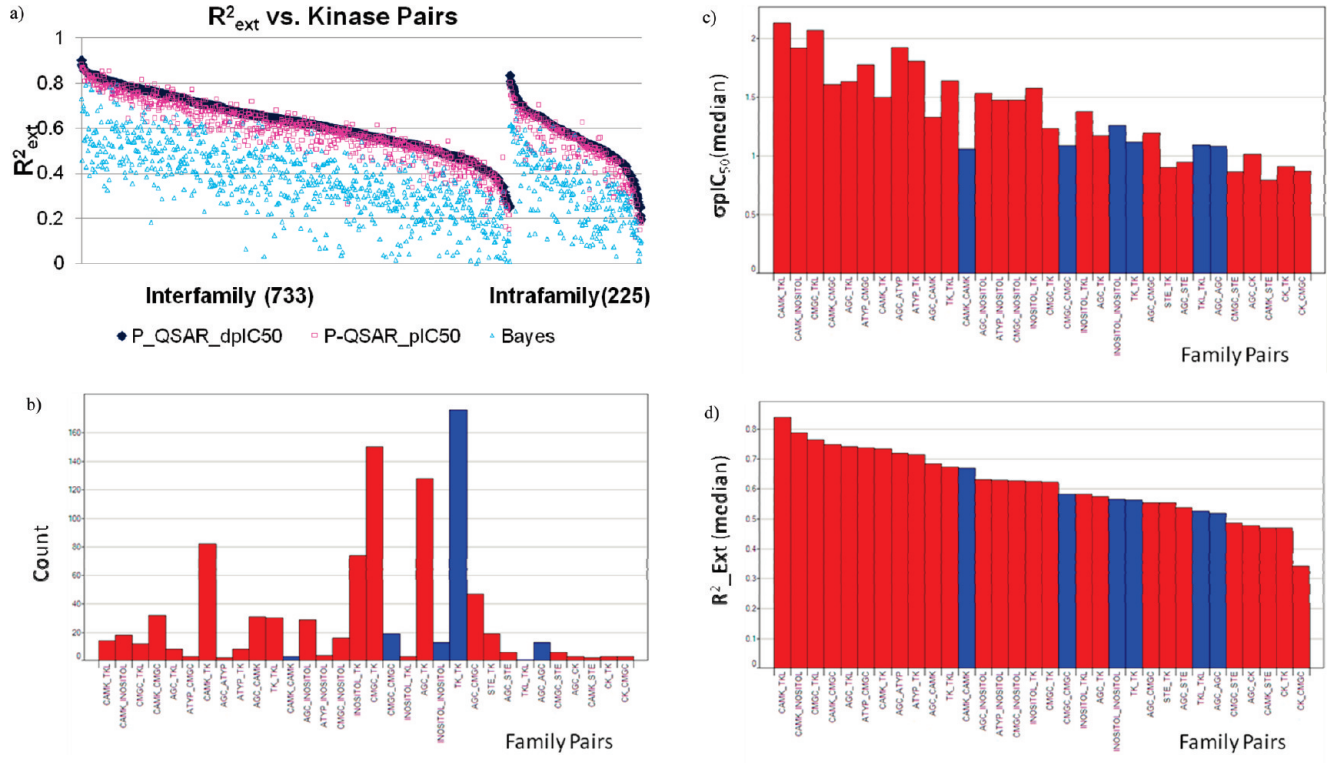


Figure 7. (a) Plot showing the R^2_{ext} obtained for the 25% held-out test sets from three methods for training Profile-QSAR selectivity models: (1) the 958 Profile-QSAR "delta models" trained directly on experimental ΔpIC_{50} as the dependent variable (in blue), (2) the 958 "difference models" from subtracting the predictions from the individual Profile-QSAR pIC_{50} models (in magenta), and (3) simple Bayesian QSAR delta models (in cyan) trained on ΔpIC_{50}. A total of 733 of these are interfamily models, while 225 are intrafamily models. (b) Histogram showing the distribution of kinase selectivity models based on the specific subfamilies of the target and antitarget kinase, sorted by decrease R^2_{ext}. Interfamily categories are in red, while intrafamily categories are in blue. (c) Bar graph showing the median dynamic range (σpIC_{50}) of data sets falling into each category. (d) Bar graph showing the median R^2_{ext} for the models belonging to each category.

958 models, the 733 interfamily selectivity models performed better that the 225 intrafamily selectivity models, with R^2_{ext} = 0.62 and $\mathbf{R}^2_{ext}$ = 0.56, respectively. There are many possible explanations, including a tendency toward slightly larger dynamic range between families (Figure 7c) and bias from projects tweaking analogues specifically to separate neighboring kinases. The distribution of kinase pairs, binned by family associations, and sorted by decreasing model quality, is shown in panel (b) of Figure 7. TK followed by CAMK and AGC families accounted for a major portion of the models. Performance for individual family pairs (Figure 7d) trends with the median dynamic range (σpIC_{50}) of the training data (Figure 7c) with an R^2 = 0.79.

Selectivity Modeling and Propagation of Error. Because the figure of merit is correlation rather than prediction residuals, the standard formula for propagating random errors in computing a difference does not apply. Numerical simulations were instead employed to evaluate how R^2_{ext} from the Profile-QSAR selectivity predictions compares to how random errors would propagate from subtracting artificially noisy pIC_{50} pairs for the same experimental sets of pIC_{50} pairs. Random Gaussian deviates were first added to the experimental pIC_{50}s in the 25% held-out test sets for the 115 assays, scaled so each artificially noisy assay reproduces the same correlation with experiment as the corresponding Profile-QSAR pIC_{50} predictions. Artificially noisy selectivities were then simulated by subtracting these synthetically noisy pIC_{50}s for each of the 958 kinase pairs. Figure 8 plots R^2 for the experimental versus Profile-QSAR predictions

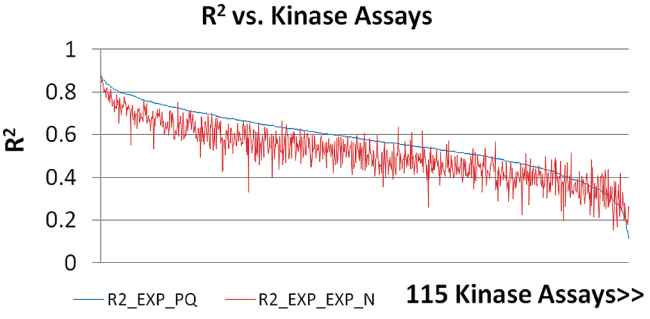


Figure 8. Plot of kinase pairs on the *X*-axis, and R^2 on the *Y*-axis. The two series, R2_EXP_PQ and R2_EXP_EXP_N, refer to the Profile-QSAR selectivity predictions vs measured experimental activity difference, and the synthetically noisy experimental activity differences vs measured experimental activity differences, respectively.

(R2_EXP_PQ) and for the experimental versus experimental + noise (R2_EXP_EXP_N), for the 958 kinase pairs. The median R^2 for the assay pairs with synthetic random noise is 0.50, significantly worse than the 0.58 for Profile-QSAR predictions reported above. For 91% of the pairs, the Profile-QSAR predictions correlate better with experiment than simulated selectivities with the same level of artificial random noise. Presumably, for some compounds, the Profile-QSAR predictions include systematic errors which deviate in the same direction for both

dx.doi.org/10.1021/ci1005004 |J. Chem. Inf. Model. 2011, 51, 1942–1956

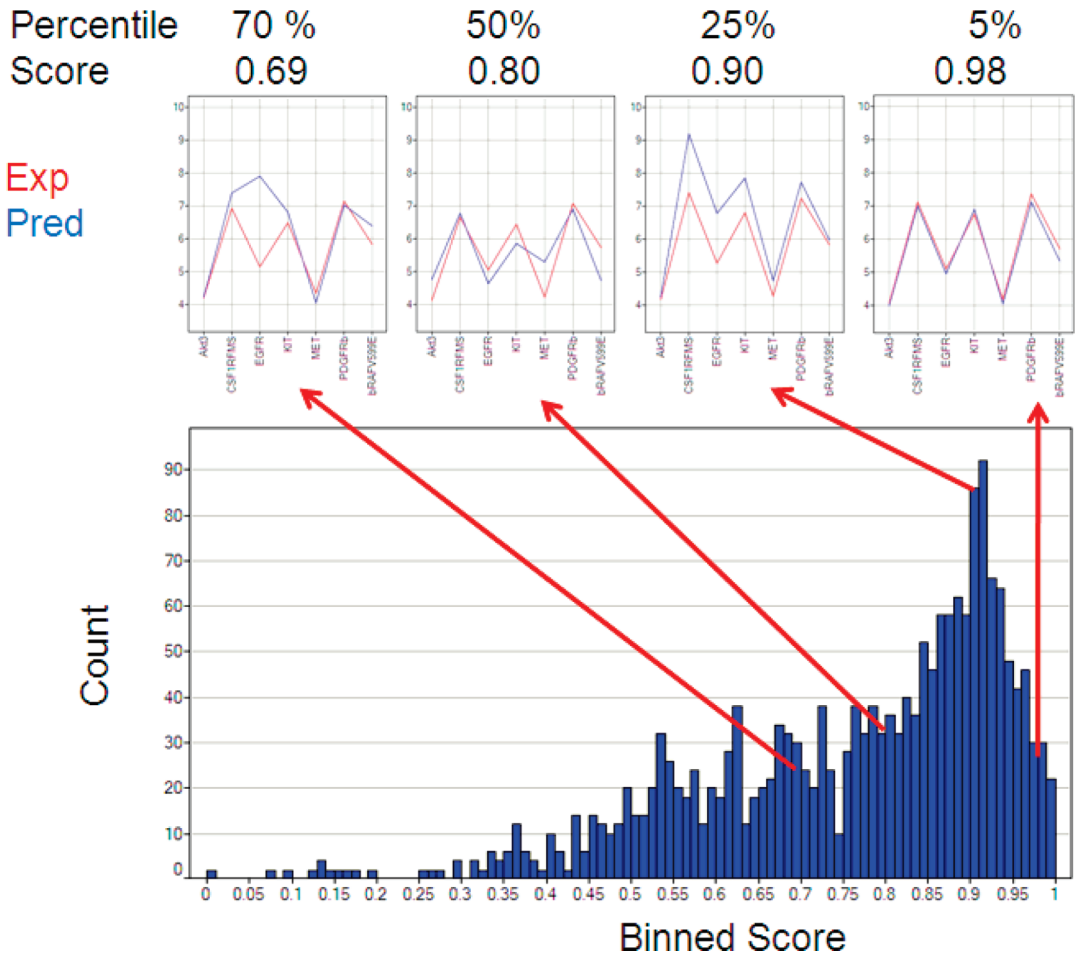


Figure 9. Histogram showing the distribution of the similarity between experimental/predicted profile pairs for 9662 compounds from 7 combinatorial libraries. The example profiles provide a visual cue on the degree of agreement between the profiles at various score thresholds.

kinases in each pair, thereby canceling and improving the selectivity predictions.

Kinase Profile Prediction. Apart from pairwise kinase selectivity, multikinase profile predictions were also evaluated, with possible applications in virtual library design and rank ordering of compounds based on their similarity to a target profile. A set of about 2700 molecules from 7 proprietary kinase-focused libraries were eliminated from the experimental activity training matrix, and the Profile-QSAR models were rebuilt using the remaining IC_{50}s. R^2 was calculated between the experimental and Profile-QSAR predictions for the 13 kinases tested against at least 100 of these compounds. Of these, nine yielded models with R^2 of at least 0.3, which was considered as the lower limit of predictive performance. The final analysis was performed on the 962 compounds that had been tested on at least 4 of the 9 kinases. The histogram in Figure 9 plots compound counts on the *Y*-axis and predicted versus experimental profile-similarity score on the *X*-axis.

$$\text{Score} = a \times R^2 + b \times E \qquad (2)$$

where

- $a \propto (\text{Experimental profile variance})^{1/5}$
- $a + b = 1$
- R^2 = Correlation squared between experimental and predicted profile.
- E = Euclidean distance between experimental and predicted profile.

The profile similarity score, given by eq 2, consists of two components: R^2 between the experimental and predicted profile, and the Euclidean distance between the experimental and predicted profile (both values are range scaled). The former measures similarity of the profile shape, while the latter indicates similarity in overall activity. The coefficients "*a*" and "*b*" are constrained to a total value of 1, and the value of "*a*" is proportional to the fifth power of the variance in the experimental profile. Thus, for most experimental profiles, with a varied shape, the score is effectively the correlation. Only for very flat profiles, where variance is meaningless, Euclidean distance, which amounts to correct average activity, determines the match. Profile similarity scores range between 0 and 1. Four example profiles illustrate the degree of agreement between the experimental and predicted profiles at various scores. The profile pairs in the top 25th percentile have excellent match, while the profile pair at the top 50th percentile shows good agreement. The example at the 70th percentile shows that six of the seven kinases are well predicted, but one is poorly predicted. Only 30% of the profiles score worse than this. Overall, the majority of profile predictions are quite acceptable.

Automated Archive Profile Prediction. An automated infrastructure was created for building models and generating full

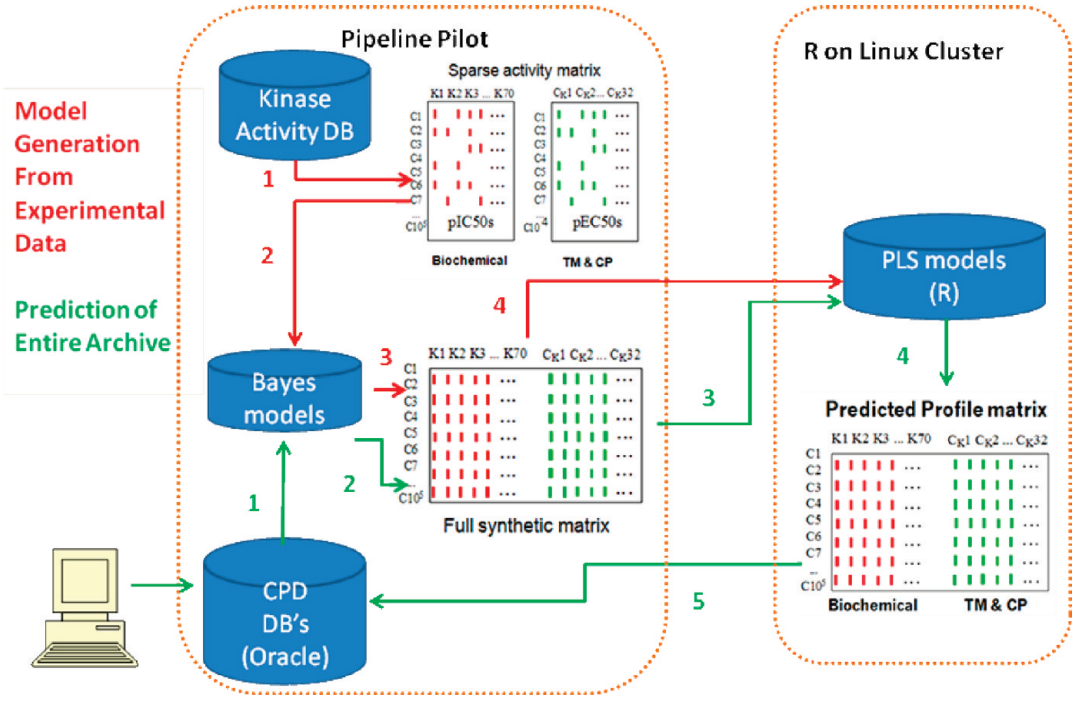


Figure 10. Workflow showing the implementation for automated model generation and profile predictions of the compound archive.

profiles across the 115 biochemical and 42 cellular assays for 2 million compounds in the corporate archive and 2 million additional drug-like commercially available compounds, using Pipeline Pilot and standalone R on a Linux cluster (Figure 10). Precomputed kinase biochemical and cellular profiles are useful filters for hit-list triaging, selectivity analysis for hit-to-lead optimizations, and library design. Additionally, as new biological data is constantly being measured, the chemical space coverage of current models increases and smaller assays accumulate enough data to qualify for model building. The red arrows represent steps for model generation, while green arrows represent steps for profile prediction. Model training (red) starts with downloading biochemical data from Oracle tables to create the sparse experimental activity matrix for 130,000 training compounds. The conventional Bayesian QSAR models are trained in Step 2, and in turn used to generate the full synthetic activity matrix in Step 3. Individual Profile-QSAR PLS models are trained in R through the batch queuing system, generating one model per core in parallel, and the resulting PLS models are stored in step 4.

The prediction phase (green arrows) starts by passing the 2 million compound corporate archive through the newly built conventional Bayesian QSARs in Step 1 to generate the corresponding synthetic activity matrix in Step 2. These are submitted (Step 3) in batches of 100,000 compounds to the R PLS models on the Linux cluster, generating full profile predictions of 115 biochemical kinase assays and 42 cellular assays for the entire corporate archive in Step 4. Depending upon the rate of addition to the activity databases, these protocols can be run on a scheduled basis to generate updated profiles.

Applications of IMTS Methodology. The results above were for retrospective applications. Since its development, IMTS with Profile-QSAR and AutoShim has also been applied prospectively in 20 therapeutic kinase projects (Figure 11a). Ten were in the initial hit-finding stages. The others aimed to identify alternative scaffolds for more advanced projects. Panel (a) of Figure 11 shows the 5-fold leave-group-out Q^2 for Profile-QSAR and

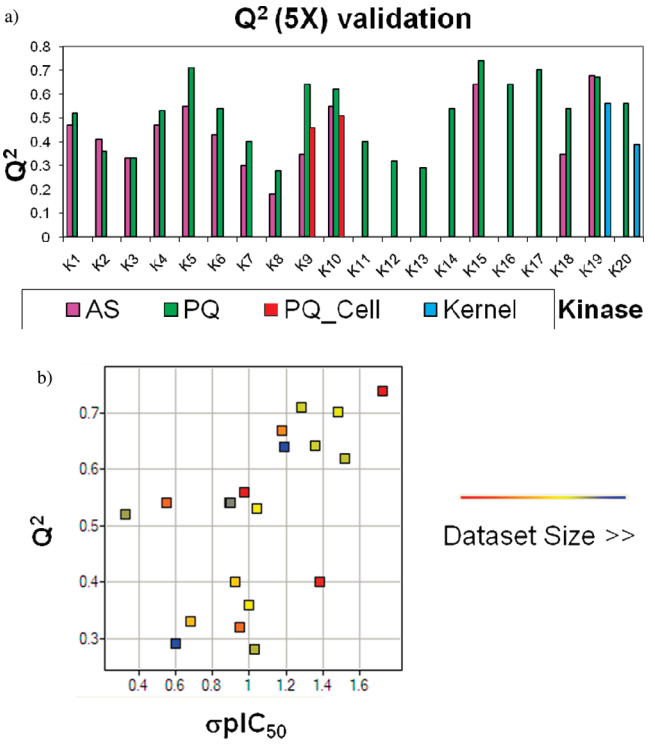


Figure 11. (a) Bar graph showing 5-fold LGO Q^2 on the *Y*-axis and the individual kinases on the *X*-axis for the Surrogate AutoShim (magenta), Profile-QSAR (green), Profile-QSAR cellular (red), and Kinase-Kernel (blue) models. (b) Plot showing Q^2 on the *Y*-axis vs standard deviation of pIC_{50} on the *X*-axis. Points are colored by data set size.

dx.doi.org/10.1021/ci1005004 |*J. Chem. Inf. Model.* 2011, 51, 1942–1956

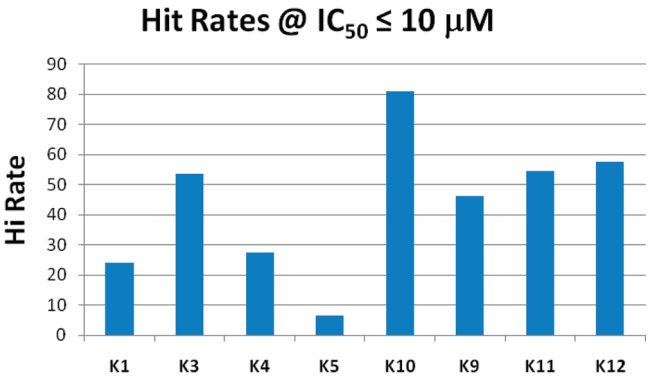


Figure 12. Bar graph showing percentage hit-rate at 10 μM for 8 kinase projects where compounds were ordered and tested. Note the no K5 compounds were predicted to reach 10 uM, so these few hits were actually slightly under-predicted.

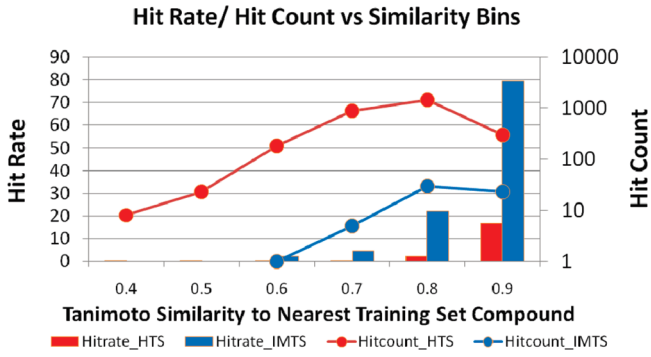


Figure 13. Histogram showing hit-rates on the first *Y*-axis, hit-count on the second *Y*-axis, and binned Tanimoto similarity of each hit to the closest neighbor from the training set on the *X*-axis. The series in red shows hit-rates from an HTS against K1, showing hit-rates from near-neighbor models. The series in blue shows hit-rates from compounds ordered based on K1 models. IMTS enrichment is an order of magnitude higher than near-neighbor modeling and recovers hits well outside of the typical SAR radius of 0.9 and higher.

AutoShim models built for these prospective applications. The Q^2 values range from ~0.3 for K7 up to ~0.7 for K5. Similar to previous reports, in 10 of 12 cases, the 2D Profile-QSAR models were slightly higher in predictive power than the 3D AutoShim target-customized docking models. As previously reported,[3] combining these two orthogonal methods routinely gave better hit recovery than either method alone. The Profile-QSAR cellular models were also employed in the K9 and K10 applications. Dynamic range and data set size correlate with predictive power (Figure 11b). K19 and K20 also used a novel kinase-kernel technique that extends Profile-QSAR predictions to kinases lacking training data, which will be the subject of an upcoming publication.

Useful enrichments were observed even for the poorest model, K3, with Q^2 only ~0.3. Note that R^2 = 0.3 is still much higher than most virtual screens by docking.[34] Figure 12 shows the percentage hit-rates at a 10 μM IC_{50} threshold for eight projects where compounds were ordered and biological results were obtained. The highest hit-rate of 80% was obtained in case of K10 followed by K12 (58%), K11 (55%), K3 (53%), K9 (46%), K4 (~28%), and K1 (~24%).

The uncharacteristically low ~7% hit-rate for K5, which had the most predictive model (R^2 > 0.7), was not a comparable study. This project's particular requirement was to find potential type II binders from a portion of the corporate archive that had been added over the most recent two years. However, there were too few experimental IC_{50}s from known type II ligands to build a specialized type II model, so a general model built from ~90% type I and ~10% type IIs had to be used. Only compounds predicted by AutoShim docking to bind in a type II pose were selected. An impressive 96% of these 12,700 compounds contained scaffolds known to bind in type II conformations, but none were predicted to be active at 10 μM. Therefore, compounds predicted to be better than 100 μM were ordered. Of these, 7% were active at 10 μM, the highest test concentration. Presumably, many more would reach the predicted 100 μM activity threshold.

K3 was an exemplary case where HTS was not possible due to difficulty with large scale protein production, but a medium-throughput single-concentration screen of ~100,000 compounds could be performed. The 1700 follow-up IC_{50}s were used to train a Profile-QSAR model, which was used to virtually screen the remaining archive.

K4 demonstrated the value of iterative model building. The first models built from the initial activity data (~2400 compounds) were not of satisfactory quality, but addition of data from a follow-up round of screening resulted in a final model with Q^2 = 0.55 and 28% hit-rate on actual ordered compounds. For K10, cellular activity predictions were factored in during the selection process.

K1's overall 19% hit-rate at 5 μM from IMTS is 27-fold higher than the 0.7% hit-rate from an earlier experimental HTS. By itself, however, this enrichment means very little. High enrichments can be achieved simply by choosing compounds similar to known actives. Anthony Nicholls has suggested near-neighbor modeling as a benchmark for model enrichment.[46] The hit-rate histogram in Figure 13 bins the compounds from the IMTS (blue) and the HTS (red) by the similarity of each compound with the nearest active training set member. Thus, the red bars correspond to the hit-rate expected from near-neighbor database searching by fishing with the known actives at that similarity threshold. The IMTS model gives almost perfect recovery in the right-most bin, whereas the hit-rate from for near-neighbor searching is about 5-fold lower within what might be considered this "SAR radius". Recovery from near-neighbor predictions in the next bin (0.8 < Tc < 0.9) drops to just 2%, whereas IMTS still recovers a respectable 22%, an 11-fold advantage. Below that similarity level, HTS hit rates fall to background, but IMTS continues to show some enrichment.

The K1 application was the very earliest, and more recent studies have shown even greater improvements in predicting novel chemistry. One likely reason for this improvement could be the addition of over 40 kinase biochemical assays to the Profile-QSAR basis set from a different site with a different history, which added many new chemotypes. In K9, models were trained on ~7000 IC_{50}s from a medium-throughput screen of 100,000 compounds. Chemical novelty was a top criterion in the selection of 319 compounds for testing, and no compounds were selected from within the SAR radius. A total of 147 compounds confirmed active at $IC_{50} \leq 10$ μM, an overall hit rate of 46%. Analogous to Figure 13, the histogram in Figure 14 shows similarity to the nearest training set compound on the *X*-axis, and hit-rate, and hit-count on the primary and secondary *Y*-axes. The series in blue

dx.doi.org/10.1021/ci1005004 |J. Chem. Inf. Model. 2011, 51, 1942–1956

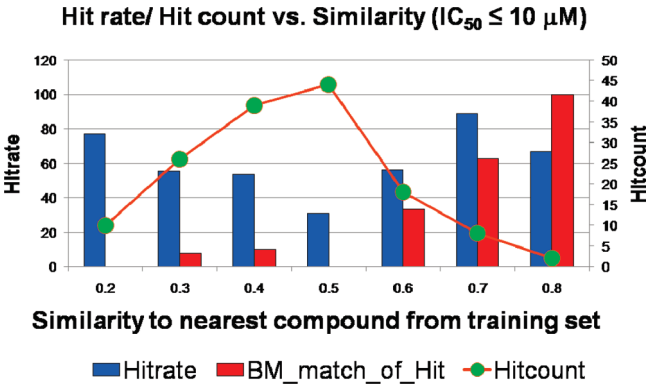


Figure 14. Analysis of hits obtained for kinase K9. Similarity to nearest active training set compound is on the *X*-axis, while the primary and secondary *Y*-axes are for hit-rate and hit-count, respectively. The three series are hit-rate (blue bars), hit-count (green circles), and percentage of hits with a Bemis and Murcko framework found in the training set (BM_match_of_hit, red bars).

bars and green circles represent the hit-rates and hit-counts, respectively. The red bars show an orthogonal novel-scaffold similarity measure, i.e., the percentage of recovered hits having a Bemis and Murcko scaffold found within the training set. It increases monotonically with Tanimoto similarity. The majority of hits were obtained at or below the 0.6 similarity bin, and only a very small percentage had previously identified scaffolds, demonstrating excellent predictive power with both novel substitutions and novel chemotypes.

DISCUSSION

Current and Envisioned Future Applications. IMTS with Profile-QSAR and AutoShim was primarily designed as an alternative to experimental HTS. It is particularly useful in projects where the target is not amenable to HTS. The K3 kinase project fell into this category, where production of protein was too difficult for full HTS. IMTS models used data from a preliminary medium-throughput screen to provide diverse hits, which were structurally novel compared to the original hits identified from the initial screening. The K1, K5, and K10 targets were mature projects where the archive had grown substantially since a prior experimental HTS had been performed. IMTS provided highly reliable virtual screens of the untested compounds added to the archive to identify potential backup scaffolds. Kinase K4 was a new target where IMTS efficiently identified chemical matter in advance of a full scale HTS. After a full experimental HTS, IMTS predictions have been used for rescuing borderline compounds and false negatives during HTS triage and to predict ligand efficiency, lipid efficiency, cellular activity, and antitarget selectivity to weight compound selection for secondary confirmation screens.

The above examples use IMTS to save time and expense of experimental HTS. Selectivity predictions and cellular potency predictions provide additional capabilities outside the realm of traditional HTS. The activity profile for 115 kinase assays and 42 cellular assays have been precomputed for the entire corporate archive and stored in the company database. Additionally, selectivity profiles are used to evaluate proposed virtual libraries and for suggesting additions to the corporate archive from commercial sources. Cellular activity predictions are particularly valuable when the cellular activity end point is affected by the inhibition of more than one kinase involving multiple pathways, and the correlation between the cellular activity end point and target of interest is low.

Future Efforts. The main limitation of IMTS is the need for initial IC_{50} training data. Kinases with high active-site sequence similarity are highly cross-reactive, although the converse in not generally true. The 92 kinases for which models now exist cover most of the kinome. A novel "Kinase-Kernel" method creates an initial model for a new kinase with no training data from a weighted average of Profile-QSAR predicted activities from neighboring kinases with similar active site sequences. This will be the subject of a forthcoming publication. Additionally, evaluations are underway for extending the methods toward the screening of fragment-like compounds. Most available training data are for drug-sized compounds that fill the active site. Adjustments are required to build Profile-QSAR and AutoShim models for smaller molecules that only occupy individual pockets. Other extensions include addition of 3D pharmacophore fingerprints and identification of predicted profile cliff pairs to enhance scaffold-hopping capability.

Initial efforts to expand Profile-QSAR and AutoShim modeling to additional protein families beyond kinases have been promising. The flexible binding site is a particular challenge for kinase modeling, but kinases do provide the advantage of a well-defined pocket with highly conserved features. Proteases have more rigid binding sites, but the pockets are shallower, with larger structural and feature diversity, and a more uneven bioactivity landscape. Covalent inhibitors provide additional challenges. IC_{50} data are also much less available for proteases than for kinases.

Nevertheless, Profile-QSAR models for 24 S1-serine proteases with at least 250 IC_{50}s and at least 15 submicromolar compounds, prefiltered for irreversible warhead groups, showed comparable predictive performance to kinases, with a median R^2_{ext} of 0.60. Interestingly, Profile-QSAR models for two Cysteine proteases, built by combining the two available Cysteine protease assays with 24 S1-Serine protease assays, showed highly improved performance over the individual Cysteine protease Bayesian-QSARs. This suggests that SAR similarity can transcend the catalytic residue-based classification of proteases.

Profile-QSAR models of one Cysteine protease identified novel scaffolds as backups to the ones then in medchem optimization. Two assays were available to build models of good predictive quality, with R^2_{ext} 0.41 and 0.56. The experimental pIC_{50}s of the target Cysteine protease were first compared to Profile-QSAR predictions for S1-serine proteases to identify potential cross-reactivity issues. In addition, 32 compounds tested from a Profile-QSAR virtual screen of the archive yielded 13 actives at the 10 μM IC_{50} threshold, a 41% hit rate. These 13 hits included three new chemotypes, currently being followed up by screening additional analogues. Details of these and other extensions to protease-family modeling will be presented in a forthcoming publication.

CONCLUSION

Experimental HTS of large corporate archives is a proven methodology for discovering chemical starting points for novel targets but carries enormous burdens of time and resources, typically six months and $1,000,000[3]. Furthermore, being tied to a preplated compound collection, HTS is not amenable to

screening external vendor collections or virtual compound libraries. The huge investment of HTS often precludes its use on nonvalidated targets to identify tool compounds that may be critical for the validation process itself. Protein targets that are hard to express are likewise not amenable to a full scale HTS. HTS can be difficult to adapt to cellular assays. Compound archives are constantly turning over, and past HTS runs quickly become incomplete. HTS results are often noisy, requiring orthogonal methods like VS for triaging the hit-lists. Additionally, Profile-QSAR's approximate IC_{50} predictions provide additional capabilities beyond single-concentration, yes/no activity predictions from experimental HTS, or other less quantitative VS methods. These include weighted sampling schemes that seek to balance predicted activity with other desirable drug-like properties, such as predicted ligand efficiency and lipophilic ligand efficiency. In silico selectivity predictions can bias selections to less promiscuous starting points. Therefore, there is a high value for computational approaches to complement early stage hit-finding initiatives, as well as catch-up screens for more mature projects.

While in silico screens are efficient and easy to implement, they must also be accurate to effectively complement traditional HTS. Docking-based VS methods rarely correlate with affinity. Conventional QSARs do better than docking but only to a limited extent. On the other hand, many pharmaceutical companies have large databases of kinase activities, and reliable medium-throughput screens are usually the first priority for early stage projects. IMTS marries the two and generates target-tailored scoring metrics trained on these experimental data.

Profile-QSAR is a novel *meta*-QSAR methodology that harnesses high quality data across targets belonging to a gene family to yield highly predictive models. It circumvents the "Achilles heel" of polypharmacology,[47,48] i.e., the sparse cross-reactivity matrix, by creating a synthetic activity matrix from simple Bayesian QSAR models to fill the 90% or more missing values. A sparse activity matrix of ~115 kinases and ~130,000 compounds containing about 1.5 million IC_{50}s (10% full) was used to generate synthetic activity values, which in turn were used as "chemical descriptors" in PLS regression to generate the *meta*-QSARs. These provide vastly superior predictive performance compared to the QSARs on which they are based, with median R^2_{ext} = 0.6 for Profile-QSAR, but only R^2_{ext} = 0.3 for Bayesian QSAR. Full profile predictions of all 92 kinases for the entire corporate collection of 2 million internal compounds, plus 2 million drug-like commercial compounds, can be generated within a few days. The method has successfully generated predictive models for 20 active kinase projects with 5X LGO Q^2 ranging from 0.3 to 0.7. Profile-QSAR predictions appear to be orthogonal to the predictions from the previously reported 3D protein-family docking-based AutoShim method, and significant boosts in the retrieval of actives when taking the union of the two predictions (observed for K1), or a linear combination of the two predictions (observed for K3 and K4), over either method alone. While both standard docking-based virtual screening and IMTS require expensive cherry-picking, docking typically gives 2-fold to 7-fold enrichments, while IMTS typically provides 20-fold to 40-fold enrichment, greatly reducing the cherry-picking selection size. Docking studies often report recovery at 10% or even 20% of the database. For a corporate archive of 2 million compounds, that would mean cherry picking 100,000s of compounds, far more expensive than just screening the whole archive. IMTS, being a data-driven method, improves with iterations of screening. Thus, by testing a smaller number of compounds, several rounds of IMTS can cover a wider range of chemical space in a more rational way and with greater flexibility than one round of selection using untrained methods such as docking. In this way, sets of only hundreds of compounds can recover large numbers of diverse novel hits.

AUTHOR INFORMATION

Corresponding Author

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

Present Addresses

†Anacor Pharmaceuticals, Inc., 1020 East Meadow Circle, Palo Alto, CA 94303

ACKNOWLEDGMENT

The authors thank Jason Kondracki, Pradeep Pasupuleti, Sandhya Sreepathy, and Joseph Ringgenberg for informatics support, and Kevin Shoemaker and Doriano Fabbro for biochemical and cellular assay support. P.M. would like to thank the NIBR Education Office for Post Doctoral funding.

REFERENCES

(1) Bleicher, K. H.; Boehm, H. J.; Mueller, K.; Alanine, A. I. A guide to drug discovery: Hit and lead generation: Beyond high-throughput screening. *Nat. Rev. Drug Discovery* **2003**, *2*, 369–378.

(2) Harris, C. J.; Stevens, A. P. Chemogenomics: Structuring the drug discovery process to gene families. *Drug Discov Today* **2006**, *11*, 880–888.

(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) Cavasotto, C. N.; Orry, A. J. W. Ligand docking and structure-based virtual screening in drug discovery. *Curr. Top. Med. Chem. (Sharjah, United Arab Emirates)* **2007**, *7*, 1006–1014.

(5) Guido, R. V. C.; Oliva, G.; Andricopulo, A. D. Virtual screening and its integration with modern drug design technologies. *Curr. Med. Chem.* **2008**, *15*, 37–46.

(6) Green, D. V. S. Virtual screening of chemical libraries for drug discovery. *Expert Opin. Drug Discovery* **2008**, *3*, 1011–1026.

(7) Kubinyi, H. Success Stories of Computer-Aided Design. In *Computer Applications in Pharmaceutical Research and Development*; Ekins, S., Ed.; WileySeries in Drug Discovery and Development; Wiley-Interscience: Hoboken, NJ, 2006; pp 377−424.

(8) Doman, T. N.; McGovern, S. L.; Witherbee, B. J.; Kasten, T. P.; Kurumbail, R.; Stallings, W. C.; Connolly, D. T.; Shoichet, B. K. Molecular docking and high-throughput screening for novel inhibitors of protein tyrosine phosphatase-1B. *J. Med. Chem.* **2002**, *45*, 2213–2221.

(9) Paiva, A. M.; Vanderwall, D. E.; Blanchard, J. S.; Kozarich, J. W.; Williamson, J. M.; Kelly, T. M. Inhibitors of dihydrodipicolinate reductase, a key enzyme of the diaminopimelate pathway of Mycobacterium tuberculosis. *Biochim. Biophys. Acta, Protein Struct. Mol. Enzymol.* **2001**, *1545*, 67–77.

(10) Birault, V.; Harris, C. J.; Le, J.; Lipkin, M.; Nerella, R.; Stevens, A. Bringing kinases into focus: Efficient drug design through the use of chemogenomic toolkits. *Curr. Med. Chem.* **2006**, *13*, 1735–1748.

(11) Oshiro, C.; Bradley, E. K.; Eksterowicz, J.; Evensen, E.; Lamb, M. L.; Lanctot, J. K.; Putta, S.; Stanton, R.; Grootenhuis, P. D. J. Performance of 3D-database molecular docking studies into homology models. *J. Med. Chem.* **2004**, *47*, 764–767.

(12) Manallack, D. T.; Pitt, W. R.; Gancia, E.; Montana, J. G.; Livingstone, D. J.; Ford, M. G.; Whitley, D. C. Selecting screening

candidates for kinase and G protein-coupled receptor targets using neural networks. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1256–1262.

(13) Xia, X.; Maliski, E. G.; Gallant, P.; Rogers, D. Classification of kinase inhibitors using a bayesian model. *J. Med. Chem.* **2004**, *47*, 4463–4470.

(14) Martin, E. J.; Sullivan, D. C. AutoShim: Empirically corrected scoring functions for quantitative docking with a crystal structure and IC_{50} training data. *J. Chem. Inf. Model.* **2008**, *48*, 861–872.

(15) Caron, P. R.; Mullican, M. D.; Mashal, R. D.; Wilson, K. P.; Su, M. S.; Murcko, M. A. Chemogenomic approaches to drug discovery. *Curr. Opin. Chem. Biol.* **2001**, *5*, 464–470.

(16) Bredel, M.; Jacoby, E. Chemogenomics: An emerging strategy for rapid target and drug discovery. *Nat. Rev. Genet.* **2004**, *5*, 262–275.

(17) Bajorath, J. Computational approaches in chemogenomics and chemical biology: Current and future impact on drug discovery. *Expert Opin. Drug Discovery* **2008**, *3*, 1371–1376.

(18) 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.

(19) Sutherland, J. J.; Higgs, R. E.; Watson, I.; Vieth, M. Chemical fragments as foundations for understanding target space and activity prediction. *J. Med. Chem.* **2008**, *51*, 2689–2700.

(20) 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.

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

(22) 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.

(23) Muegge, I.; Enyedy, I. J. Virtual screening for kinase targets. *Curr. Med. Chem.* **2004**, *11*, 693–707.

(24) Sheridan, R. P.; Nam, K.; Maiorov, V. N.; McMasters, D. R.; Cornell, W. D. QSAR models for predicting the similarity in binding profiles for pairs of protein kinases and the variation of models between experimental data sets. *J. Chem. Inf. Model.* **2009**, *49*, 1974–1985.

(25) Aronov, A. M.; McClain, B.; Stuver Moody, C.; Murcko, M. A. Kinase-likeness and kinase-privileged fragments: Toward virtual polypharmacology. *J. Med. Chem.* **2008**, *51*, 1214–1222.

(26) Fernandez, A.; Maddipati, S. A priori inference of cross reactivity for drug-targeted kinases. *J. Med. Chem.* **2006**, *49*, 3092–3100.

(27) Posy, S. L.; Hermsmeier, M. A.; Vaccaro, W.; Ott, K. H.; Todderud, G.; Lippy, J. S.; Trainor, G. L.; Loughney, D. A.; Johnson, S. R. Trends in kinase selectivity: Insights for Target class-focused library screening. *J. Med. Chem.* **2011**, *54*, 54–66.

(28) Bajorath, J. Affinity fingerprints: Leading the way? *Drug Discov Today* **2002**, *7*, 1035

(29) Dixon, S. L.; Villar, H. O. Bioactive diversity and screening library selection via affinity fingerprinting. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 1192–1203.

(30) Beroza, P.; Damodaran, K.; Lum, R. T. Target-related affinity profiling: Telik's lead discovery technology. *Curr. Top. Med. Chem. (Sharjah, United Arab Emirates)* **2005**, *5*, 371–381.

(31) Kauvar, L. M.; Higgins, D. L.; Villar, H. O.; Sportsman, J. R.; Engqvist-Goldstein, A.; Bukar, R.; Bauer, K. E.; Dilley, H.; Rocke, D. M. Predicting ligand binding to proteins by affinity fingerprinting. *Chem. Biol.* **1995**, *2*, 107–118.

(32) Bender, A.; Jenkins, J. L.; Glick, M.; Deng, Z.; Nettles, J. H.; Davies, J. W. "Bayes Affinity Fingerprints" Improve retrieval rates in virtual screening and define orthogonal bioactivity space: When are multitarget drugs a feasible concept? *J Chem. Inf. Model.* **2006**, *46*, 2445–2456.

(33) Lessel, U. F.; Briem, H. Flexsim-X: A method for the detection of molecules with similar biological activity. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 246–253.

(34) 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.

(35) Manning, G.; Whyte, D. B.; Martinez, R.; Hunter, T.; Sudarsanam, S. The protein kinase complement of the human genome. *Science* **2002**, *298*, 1912–1916. 1933.

(36) Naive Bayes classifier. http://en.wikipedia.org/wiki/Naive_Bayes_classifier (accessed March 1, 2011).

(37) The R Project for Statistical Computing. www.r-project.org (accessed March 1, 2011.

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

(39) *Pipeline Pilot 8.0 Data Modeling User Guide*; Accelrys, Inc.: San Diego, CA , 2011.

(40) Agarwal, A.; Pearson, P. P.; Taylor, E. W.; Li, H. B.; Dahlgren, T.; Herslof, M.; Yang, Y.; Lambert, G.; Nelson, D. L. Three-dimensional quantitative structure-activity relationships of 5-HT receptor binding data for tetrahydropyridinylindole derivatives: A comparison of the Hansch and CoMFA methods. *J. Med. Chem.* **1993**, *36*, 4006–4014.

(41) Thomas, B. F.; Compton, D. R.; Martin, B. R.; Semus, S. F. Modeling the cannabinoid receptor: A three-dimensional quantitative structure-activity analysis. *Mol. Pharmacol.* **1991**, *40*, 656–665.

(42) Goodwin, J. T.; Conradi, R. A.; Ho, N. F. H.; Burton, P. S. Physicochemical determinants of passive membrane permeability: Role of solute hydrogen-bonding potential and volume. *J. Med. Chem.* **2001**, *44*, 3721–3729.

(43) Murphy, E. A.; Shields, D. J.; Stoletov, K.; Dneprovskaia, E.; McElroy, M.; Greenberg, J., I; Lindquist, J.; Acevedo, L. M.; Anand, S.; Majeti, B. K.; Tsigelny, I.; Saldanha, A.; Walsh, B.; Hoffman, R. M.; Bouvet, M.; Klemke, R. L.; Vogt, P. K.; Arnold, L.; Wrasidlo, W.; Cheresh, D. A. Disruption of angiogenesis and tumor growth with an orally active drug that stabilizes the inactive state of PDGFRbeta/B-RAF. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107*, 4299–4304.

(44) Nazarian, R.; Shi, H.; Wang, Q.; Kong, X.; Koya, R. C.; Lee, H.; Chen, Z.; Lee, M. K.; Attar, N.; Sazegar, H.; Chodon, T.; Nelson, S. F.; McArthur, G.; Sosman, J. A.; Ribas, A.; Lo, R. S. Melanomas acquire resistance to B-RAF(V600E) inhibition by RTK or N-RAS upregulation. *Nature (London, U. K.)* **2010**, *468*, 973–977.

(45) Solit, D. B.; Rosen, N. Resistance to BRAF inhibition in melanomas. *N. Engl. J. Med.* **2011**, *364*, 772–774.

(46) Nicholls, A. *Information Theory and QSAR*. 18th European Symposium on Quantitative Structure Activity Relationships, Rhodes, Greece, September 19−24, 2010.

(47) Mestres, J.; Gregori-Puigjane, E. Conciliating binding efficiency and polypharmacology. *Trends Pharmacol. Sci.* **2009**, *30*, 470–474.

(48) Keiser, M. J.; Roth, B. L.; Armbruster, B. N.; Ernsberger, P.; Irwin, J. J.; Shoichet, B. K. Relating protein pharmacology by ligand chemistry. *Nat. Biotechnol.* **2007**, *25*, 197–206.