JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

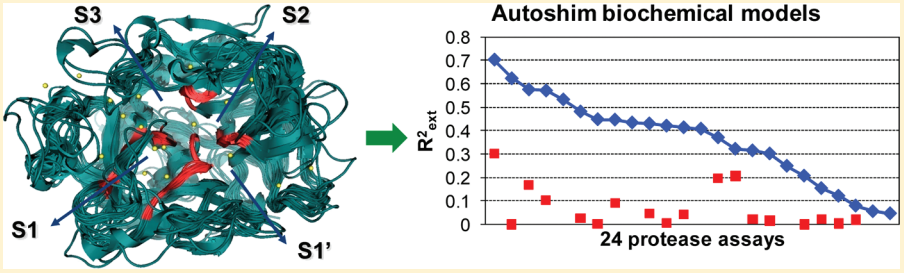
pubs.acs.org/jcim

Profile-QSAR and Surrogate AutoShim Protein-Family Modeling of Proteases

Prasenjit Mukherjee*,† and Eric Martin

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

Supporting Information



ABSTRACT: The 2D Profile-QSAR and 3D Surrogate AutoShim protein-family virtual screening methods were originally developed for kinases. They are the key components of an iterative medium-throughput screening alternative to expensive and time-consuming experimental high-throughput screening. Encouraged by the success with kinases, the S1-serine proteases were selected as a second protein family to tackle, based on the structural and SAR similarity among them, availability of structural and bioactivity data, and the current and future small-molecule drug discovery interest. A validation study on 24 S1-serine protease assay data sets from 16 unique proteases gave very promising results. Profile-QSAR gave a median R_{ext}^2 = 0.60 for 24 assay data sets, and pairwise selectivity modeling on 60 protease pairs gave a median R_{ext}^2 = 0.64, comparable to the performance for kinases. A 17-structure universal ensemble S1-serine protease surrogate receptor for Autoshim was developed from a collection of ∼1500 X-ray structures. The predictive performance on 24 S1-serine protease assays was good, with a median R_{ext}^2 = 0.41, but lower than had been obtained for kinases. Analysis suggested that the higher structural diversity of the protease structures, as well as smaller assay data sets and fewer potent compounds, both contributed to the decreased predictive power. In a prospective virtual screening application, 32 compounds were ordered from a 1.5 million archive and tested in a biochemical assay. Thirteen of the 32 compounds were active at $IC_{50} \leq 10$ μM, a 41% hit-rate. Three new scaffolds were identified which are being followed up with testing of additional analogues. A SAR similarity analysis for this target against 13 other proteases also indicated two potential protease targets which were positively and negatively correlated with the activity of the target protease.

INTRODUCTION

2D Profile-QSAR[1] and 3D Surrogate AutoShim[2,3] are the foundations of an internally developed iterative medium-throughput screening (IMTS) methodology for virtual screening (VS) compound archives against kinases, as an alternative to expensive, time-consuming, experimental high-throughput screening (HTS). Profile-QSAR is a "meta-QSAR" method which 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 of 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 (MTS) data, for predicting activity against new kinases. Along with biochemical IC_{50} predictions, the method predicts pairwise kinase selectivity, kinase biochemical profiles, and cellular activity of kinase inhibitors.

The 3D Surrogate AutoShim method uses several hundred experimental IC_{50}s to "shim" a "universal ensemble" of surrogate X-ray crystal structures, creating a target-tailored scoring function that predicts affinity much more accurately than conventional docking. Furthermore, the same surrogate ensemble is used for the entire protein family, i.e., kinases, so the entire corporate archive plus 2 million drug-like commercial compounds were predocked just one time. The protein−ligand pharmacophore interactions were extracted and stored in a database. New pharmacophore shims can quickly be trained and scored for each new target in the family, delivering accurate activity predictions in a day, rather than weeks or months for conventional docking.

The methods have been successfully applied to >20 kinase projects in a wide variety of roles: target tailored virtual

Received: October 23, 2011
Published: May 21, 2012

dx.doi.org/10.1021/ci300059d | J. Chem. Inf. Model. 2012, 52, 2430−2440

screening, HTS-triaging, isoform selectivity profiling, library design and enhancement, SAR similarity analysis, and understanding complex cell proliferation end-points through multi-kinase inhibition characteristics. The median hit-rate at $IC_{50} \leq 10\ \mu M$ for 9 virtual screening applications was 53%. Furthermore, these compound selections have emphasized chemical novelty, and the high hit-rates consistently include both substituents and scaffolds not found in the training sets. A HTS triaging experiment using in silico profiling against a large kinase panel helped select a set of 20 compounds for full experimental profiling from an MTS of 100 000 compounds. The best candidate had $IC_{50} = 0.05\ \mu M$ against the target of interest, 1.2−10 μM IC_{50} for 8 antitargets, and >10 μM IC_{50} for 62 other antitargets. Analysis of cell proliferation data of Raf inhibitors using Profile-QSAR helped identify PDGFRb as a potential target sharing polypharmacology with Raf kinase, and their interplay was subsequently observed in clinical studies of a Raf inhibitor.[4−6] A publication summarizing some kinase success stories is currently in preparation.

Several factors made kinases[7] the obvious first-choice for development of protein-family based models. Kinases display significant polypharmacology, i.e., most inhibitors hit multiple kinases. Traditional kinase inhibitors bind in the ATP site, which shares many common interactions, allowing a sampling of superimposed structures to serve as a universal surrogate ensemble receptor for AutoShim modeling. The common interactions form the basis for modeling new kinases as appropriately weighted hybrids of previously studied family members. Novartis, a pioneer of kinase drug discovery, has a wealth of kinase crystallographic and bioactivity data for robust model building. Current and future research interests in this area are high and have fueled the large number of applications over a 3 y period. Expanding the techniques to additional protein families would have to meet similar requirements.

Proteases[8−10] catalyze the hydrolysis of peptide bonds. Endopeptidases cleave in the middle, and exopeptidases cleave terminal residues (Amino peptidases cleave at the N terminus, while carboxy peptidase cleaves at the C terminus). On the basis of their catalytic mechanism and their main catalytic residue, they may be classified as Aspartic, Glutamic (not found in mammals), Metallo, Serine, or Cysteine proteases. They play a key role in physiological processes such as hemostasis (coagulation), tissue remodeling, wound healing, immune response, cell-cycle progression, cell proliferation/cell death, and DNA replication. They have provided a wide range of drug targets.[11−13] Some have produced successful drugs: Ace[14] (Metallo), HIV[15] protease (Aspartic), Thrombin (Serine),[16] Factor Xa[17] (Serine). Others, like matrix metalloprotease, have proved challenging.[18] Small molecule inhibitors for proteases can act via covalent and noncovalent mechanism. Covalent inhibitors[19] form a chemical bond with the catalytic residue to permanently inactivate the enzyme. Noncovalent inhibitors form a nonbonded complex with the enzyme, reversibly inhibiting its catalytic function. Certain inhibitor chemotypes display specific pharmacophoric features, e.g. metal chelating groups on Metalloprotease inhibitors, which are critical to their activity.

The choice of using S1-serine proteases[20] for the first protease extension of Profile-QSAR and AutoShim was governed by target knowledge, data availability, and use for current and future drug development. Among the four groups of proteases classified based on their catalytic mechanism: Aspartic, Cysteine, Metallo, and Serine, the last two are the most abundant in vertebrates. While Aspartic and Metallo proteases use distinct general-base mechanisms, both Serine and Cysteine protease use similar catalytic triad based mechanisms for substrate processing. Within Serine proteases, the S1 subfamily provides a rich diversity of targets and has received the highest research interest. Consequently, this subfamily has the most bioactivity and structural data. Furthermore, the S1-serine protease subfamily has shown significant polypharmacology, as reported in literature and seen in the SAR analysis below. They share similar binding sites and overall protein domains, providing a structural basis for ensemble modeling. Other serine protease subfamilies, e.g. DPP4 and HCV protease, are distinct from the S1-serine proteases, and cannot easily be modeled as the same protein-family. Certain cysteine proteases, however, do share similarities with the S1-serine proteases, e.g. Cathepsin K may be included in the same group.

METHODS

Resources and Software. Pipeline Pilot [Accelrys, San Diego, CA] was used for all data preparation, Bayesian-QSAR model building, and "synthetic activity" matrix generation for Profile-QSAR and for selection of compounds for virtual screening and analysis of the cysteine protease "C1". The PLS step of Profile-QSAR and PLS-RP step for AutoShim used R [www.r-project.org] scripts. For AutoShim model building, scripts from the kinase implementation were reconfigured and used. New shell wrappers were written for the self-docking pose-validations. Analysis was done in Tibco Spotfire and Microsoft Excel. Protein structures were aligned in MOE [chemical computing group, Montreal, Canada]. The Pprep utility from Schrödinger [Portland, OR] was used for molecular mechanics refinement of protein structures. Docked poses generated with Dockit [Metaphorics LLC, Aliso Viejo, CA] were minimized and scored with Flo+ [Boston De Novo, Boston, MA]. 3D pharmacophore descriptors were calculated with Magnet [Metaphorics LLC, Aliso Viejo, CA].

Data Preparation. Biochemical data for Profile-QSAR and AutoShim model building were downloaded from an internal database with an internal interfacing tool. Activities beyond the upper or lower limits of the assays range were offset by a factor of 10 in the appropriate direction to include information from the inactive and extremely active compounds. Pipeline Pilot filters removed compounds containing known irreversible substructures. The activity data (IC_{50}) was converted to the linear scale (pIC_{50}) by doing the log transformation

$$pIC_{50} = -\mathrm{Log}_{10}(IC_{50} \times 10^{-6}) \quad (1)$$

SAR Similarity. A SAR similarity analysis was carried out using 24 S1-serine protease assays. All unique pairs of proteases with >100 compounds tested in both assays were selected. If multiple assays existed for the same protease, only the assay with the largest intersection with the other enzyme in the pair was employed. SAR similarity was calculated as correlation (R) between the pIC_{50}s from the paired protease (or kinase) assays, while sequence similarity was the whole protease (or kinase) domain percent sequence identity calculated following multi-sequence alignment.

Profile-QSAR. Similar to previously reported kinase work, Profile-QSAR model building commenced with the generation of a complete "synthetic activity" matrix using Pipeline Pilot's Bayesian-QSAR.[21,22] As a compromise between binary (active/

inactive) and continuous modeling, up to four Bayesian categorization models were built for each kinase, using up to four concentration thresholds of 5, 10, 20, and 40% of the pIC_{50} distribution for each kinase, provided that the active category had ≥25 members. This varies slightly from the kinase implementation, where up to 5 concentration thresholds were used at 0.1, 0.3, 1, 3 and 10%. The protease sets on average are much smaller than the kinase sets, and the original parameters led to very few models using the higher thresholds. The molecular descriptors used for the Bayesian QSAR were FCFP_6 (functional class extended-connectivity fingerprint of maximum diameter 6) fingerprints plus 5 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. The 24 S1-serine protease assays produced 78 bayesian prediction columns in the synthetic activity matrix. For the extended validation with cysteine protease data sets, the same overall workflow was used with 10 additional cysteine protease assays providing 36 additional bayesian prediction columns.

The panel of Bayesian models was used to generate predictions on all unique compounds in all the activity data sets involved in model building. This resulted in a full synthetic matrix of compounds by prediction of Bayesian probabilities. As with the kinases, the predicted Bayesian probabilities in the synthetic activity matrix for all the protease assays were then used as chemical descriptors in PLS regression of each new protease for the Profile-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_{scaled}^2" was defined to penalize models with many LVs:

$$Q_{scaled}^2 = Q^2 - (0.002LV) \quad (2)$$

Up to 25 LVs were allowed, albeit rarely needed, and the model with the highest Q_{scaled}^2 was selected for prediction on the 25% held-out external test set. R_{ext}^2 from the predictions on the test set was used for final model quality assessment. As described for kinases[1] R_{ext}^2, i.e. correlation squared between experimental and predicted pIC_{50} on the held-out test set, was used as the measure of predictive performance.

For Profile-QSAR selectivity modeling on S1-serine proteases, both "delta models" and "difference models" were built. All possible protease pairs from the 24 protease assays passing a minimum data set criterion of ≥250 common IC_{50}s and ≥15 compounds with ≥3 log order activity difference were selected. A 25% held-out test set was created for each pair from the common data points. For "delta" models, Profile-QSAR models for selectivity ($dpIC_{50}$) were trained directly on the remaining 75% of *experimental* pIC_{50} pair activity differences. For the "difference models", all the remaining data in the individual pIC_{50} data sets were used to generate Profile-QSAR models for the individual proteases, and $dpIC_{50}$ was calculated from the difference in *predicted* pIC_{50}s.

AutoShim. Structures preselected for pose-validation studies were prepared using Pprep. Appropriate protomeric, tautomeric, and rotameric states were assigned and incorrect atom/bond type assignments were corrected for the ligand−protein complex. Restrained minimization using the OPLS 2005 force field relieved clashes and strained geometries in the binding site. The preselected structures were aligned in MOE based on binding site residues. An additional set of residue constraints were defined in the binding site region of the multisequence alignment by considering conserved residue patterns and position of the C_α residues based on whole protease domain alignment. The structures were realigned using the set of constrained binding site residues to generate the final ensemble for Surrogate AutoShim.

Appropriate protein, ligand, and binding site definition files were created for Dockit and Flo+ using command line utilities and the Lazymouse GUI, respectively. An automated shell wrapper read in the input files and carried out the docking and minimization of poses similar to the kinase AutoShim protocol. The output was a list of in-place heavy atom RMSDs between the generated poses with the cognate ligand pose at the end of Pprep minimization calculated using a python script available from Schrodinger's script center. Structures passing the pose-validation study were used for the surrogate ensemble.

To optimize the position and threshold of the pharmacophoric shims, up to 9 geometrically diverse poses were selected for 2000 molecules which were docked into the protease ensemble based on the procedure described above. These ligands are protease inhibitors with an $IC_{50} < 10\ \mu M$ against at least one of the proteases associated with the validation data sets. The coordinates of the polar, i.e. N, O, and nonpolar, i.e. C, F, Cl, Br, ligand atoms from the geometrically diverse poses were extracted and a full linkage clustering routine written in R was utilized to cluster the points and select cluster representatives which would sample from areas of higher density of the specific type of atoms within the ensemble's binding site. Clustering at various thresholds was utilized to generate 20, 40, 60, and 80 cluster centers for the polar and nonpolar atoms. The polar and nonpolar cluster centers were then utilized for defining the polar and nonpolar pharmacophore features respectively. Three different pharmacophore radii of 1.5, 2.5, and 3.5 Å were evaluated and for a given set of pharmacophore definitions a specific threshold was used for defining all the polar and nonpolar features. The polar features included hydrogen bond acceptors and hydrogen bond donors while the nonpolar features included heavy atom count, aromatic atom count, and hydrophobic atom count. For a polar feature, a spherical region centered on a given cluster center and a radius equivalent to the given threshold was defined. For evaluation, the distance from the center of the feature to the closest hydrogen bond donor/acceptor of the ligand pose located within the spherical region was recorded. If such a ligand feature exists, then the feature would evaluate to a value ranging from 0 to the threshold distance. In its absence, the feature value was set to 1000. For the nonpolar features, spherical regions were defined in the same manner and number of ligand atoms of a specific kind i.e. hydrophobic, aromatic, etc., from the ligand pose, located within that spherical region was recorded. The pharmacophore implementation was carried out using the Magnet program which uses a specific file with .sea extension for capturing the pharmacophore definitions. Since multiple sets of pharmacophore definitions had to be evaluated, a C program wrapped in a shell script was utilized which took a list of coordinates and a list of pharmacophore thresholds and automatically generated multiple pharmacophore definition (.sea) files creating all possible combinations

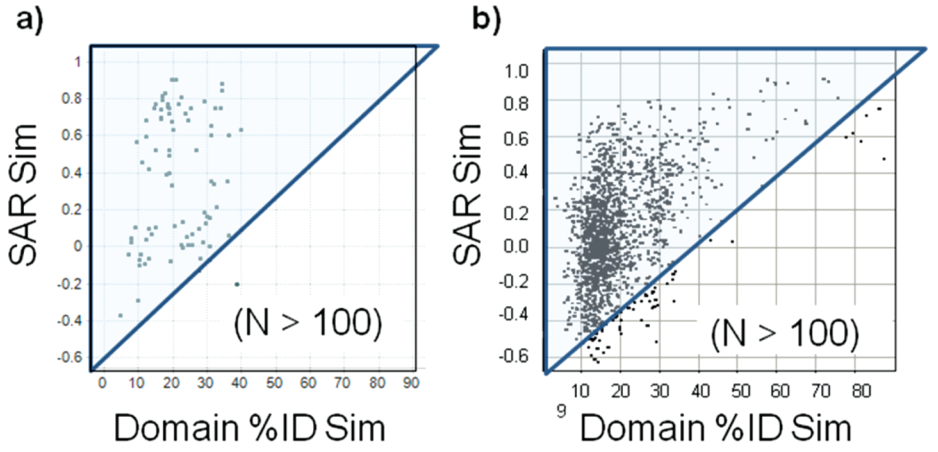


Figure 1. Plots showing SAR similarity on the *Y*-axis and percent sequence identity on the *X*-axis for pairs of (a) proteases and (b) kinases, with each pair having more than 100 common compounds tested. SAR similarity was calculated as correlation (*R*) between the pIC_{50}s from the kinase assays pairs. Domain %ID Sim was the whole protease or kinase domain percent sequence identity calculated following multisequence alignment. The tinted triangles highlight the regions of the plots corresponding to local, but not global cross-reactivity.

of the feature point definitions and thresholds for evaluation. The Magnet program uses SMART patterns to define the various forms of ligand features such as acceptors, donors, etc.

The AutoShim model building procedure has been described in detail in our previous publications on the kinase specific implementation. However, for continuity we have provided a brief description of the process here. The PLS-RP procedure coded in R handles the AutoShim model building. Each iteration of model building starts by extracting the docking score, pharmacophore values, and a few physicochemical properties for the best pose from each training set ligand based on the current docking score (initially just Flo+). Recursive partitioning over these variables is performed on the training set at six different activity thresholds with a maximum tree depth of seven. The combination of shims from the nodes of each tree define the cross-terms, or "multipoint shims". Single-point shims are taken from the tree at the single threshold of $pIC_{50} = 5.8$. The recursive partitioning defines not only which cross-terms to include, but also the thresholds, i.e. the number of non-polar or aromatic atoms in a pharmacophore or the cutoff distance to count an H-bonding feature. While the pharmacophore features are continuous descriptors, a "shim" is a categorical descriptor evaluating to a value of 1 or 0 based on the agreement/disagreement with the given evaluation statement. As an example, if we have two pharmacophore features HA1 and HD1, an example of a single point shim could be the conditional statement HA1 < 3 and a multipoint shim could be the conditional statement HA1 < 3 and HD1 < 2.5. For a given pose where HA1 = 2.9 and HD1 = 3.1, the single-point shim HA1 < 3 will evaluate to 1 since the given condition is met while the multipoint shim HA1 < 3 and HD1 < 2.5 will evaluate to 0 as the latter of the two conditions was not met.

The single- and multipoint shims are then evaluated for all the training set ligands and a partial-least squares model is trained in R on the pIC_{50} data of the ligands using the docking score and single- and multipoint shims as the independent variables. The model in turn is used to predict IC_{50} for up to nine poses per ligand, and the best scoring pose for each ligand based on the model predictions is selected for the next iteration. While the single-point shims are regenerated at each iteration, the multipoint shims from previous iterations are carried forward while new multipoint shims are added to the descriptor matrix. Successive iterations are continued, and the model building converges quickly, in 6−7 iterations, to a set of self-consistent poses in agreement with a global pharmacophoric hypothesis and a robust predictive model. Similar to Profile-QSAR, up to 25 LV's are evaluated at each iteration of model building and a process identical to Profile-QSAR (described above) is used for selecting the appropriate number of LV's for choosing the final model for a given iteration.

Chemical Novelty. For each predicted active, similarity was calculated to the nearest active training set compound with experimental $pIC_{50} \geq 5$. Similarly, each hit was decomposed to its Bemis and Murcko scaffold[23,24] and compared against the Bemis and Murcko scaffolds represented in the training set compounds with experimental $pIC_{50} \geq 5$.

RESULTS

Data Set Preparation. A collection of 17 known irreversible covalent inhibitor substructures was generated based on literature sources.[19] Filtering with these substructures, the 8596 compounds evaluated in 24 S1-serine protease assays from 16 unique enzymes identified 104 potential irreversible inhibitors, which were removed from further calculations. All the 24 data sets still conformed to the minimum data set requirements for modeling, i.e. ≥ 250 data points and ≥15 submicromolar compounds. The final data sets ranged in size from 223 to 3366 pIC_{50}s, with σ pIC_{50} ranging from 0.92 to 2.21.

SAR Similarity. Among the 24 S1-serine protease assays were 84 unique pairs of proteases with at least 100 compounds tested against both targets. Figure 1a plots the SAR similarity for each pair on the *Y*-axis, and the whole protease domain sequence identity on the *X*-axis. Figure 1b shows a corresponding plot for kinases. For kinases, all points lie roughly in the upper-left blue-tinted triangle, which highlights the region of the plot corresponding to local, but not global, cross-reactivity, i.e., when sequence identity is high, SAR similarity is generally high as well. This is shown by the points on the top right corner of the plot in Figure 1b, and the complete absence of points in the lower-right corner of the plot. However, the converse was not true—low sequence similarity can correspond to either low or high SAR similarity as

dx.doi.org/10.1021/ci300059d | J. Chem. Inf. Model. 2012, 52, 2430−2440

shown by the vertical scatter of points in the left half of Figure 1b.

Figure 1a shows that the points also lie in the upper-left triangle for proteases. As in kinases, low protease sequence similarity can correspond to either high or low SAR similarity. However, the cross-reactivity among proteases with high whole domain sequence identity cannot be determined, since no such homologous protease pairs existed with sufficient assay data. While no individual pairs have high SAR similarity, Profile-QSAR harnesses multicolinearities to improve upon the predictive performance of models trained only on data from the single protein target of interest. The overall correlation between SAR similarity and whole kinase domain sequence identity was very low, with $R^2 = 0.22$. However, the correlation between SAR similarity and whole protease domain sequence identity is nonexistent, with $R^2 = 0.01$.

Profile-QSAR. Figure 2 shows the predictive performance of Profile-QSAR for 24 S1-serine protease assay data sets. The

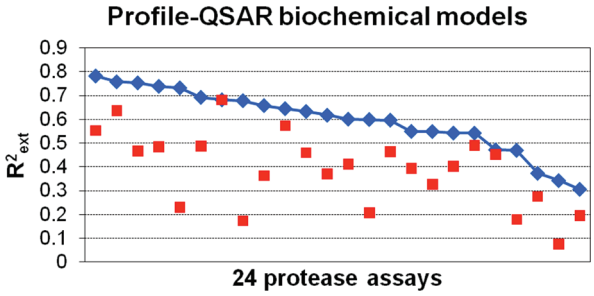


Figure 2. Plot showing the results of Profile-QSAR biochemical modeling. R_{ext}^2 is shown on the *Y*-axis for 24 protease assays shown on the *X*-axis. The performance of Profile-QSAR models are shown as blue diamonds, with a median $R_{ext}^2 = 0.6$. The corresponding Bayesian-QSAR models are shown as red squares, with median $R_{ext}^2 = 0.40$.

Profile-QSAR models are shown as blue diamonds, with median $R_{ext}^2 = 0.6$. The corresponding Bayesian-QSAR models are shown as red squares, with median $R_{ext}^2 = 0.40$. Past experience with Profile-QSAR on kinases has shown that models with $R_{ext}^2 \sim 0.3$ gives 20−40 fold enrichment of actives and have been successfully applied in dozens of project applications. The R_{ext}^2 for the 24 S1-serine protease data sets ranged from 0.7 to 0.3, generally above the minimum threshold of 0.3 required for prospective applications. On average, there is a 0.2 R^2 unit improvement in predictive power going from Bayesian QSAR to Profile-QSAR while individual improvements range from 0−0.5 to R^2 units. The dramatic improvement in predictive power is due to the vast amount of experimental data that contributes to every prediction. Through the 2-step process of first combining the data within each protease through the individual Bayesian-QSAR models, then combining it across the proteases through the final PLS equation, all 20 610 experimental IC_{50}s and 8492 chemical structures contribute to every protease activity prediction. This is far more training data than in any individual QSAR model.

Here, 60 selectivity data sets could be constructed from the 24 S1-serine protease assay data sets to evaluate the performance of Profile-QSAR in modeling selectivity (Figure 3). Each selectivity data set represents a unique protease pair in which ≥250 compounds have been tested in both the proteases, and there are ≥15 compounds with ≥1000× activity difference. As mentioned in the Methods section, selectivity

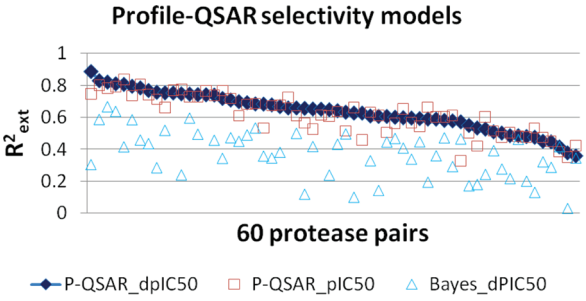


Figure 3. Results of Profile-QSAR selectivity modeling. R_{ext}^2 is shown on the *Y*-axis, while the 60 protease pairs are listed on the *X*-axis. Three series types of calculations are represented: (1) The dark blue diamonds show Profile-QSAR_dpIC$_{50}$ "delta models" trained directly on experimental dpIC$_{50}$ as the dependent variable, with median $R_{ext}^2 = 0.64$. (2) The red squares show Profile-QSAR_pIC$_{50}$ "difference models" from subtracting the predicted pIC$_{50}$s from the individual Profile-QSAR models with median $R_{ext}^2 = 0.61$. (3) The cyan triangles show simple Bayesian-QSAR models trained directly on experimental dpIC$_{50}$ (Bayes_dpIC$_{50}$) as the dependent variable with much lower median $R_{ext}^2 = 0.39$.

was predicted both by subtracting predicted pIC$_{50}$s from the individual biochemical activity Profile-QSAR_pIC$_{50}$ models (difference models) and by training 60 new Profile-QSAR_dpIC$_{50}$ models directly on the difference in experimental pIC$_{50}$ (dpIC$_{50}$) as the dependent variable (delta models). Profile-QSAR_dpIC$_{50}$ models trained directly on experimental dpIC$_{50}$, with a median $R_{ext}^2 = 0.64$ performed slightly better than simply subtracting predicted pIC$_{50}$s, which had a median $R_{ext}^2 = 0.61$. The correlation for corresponding Bayesian_dpIC$_{50}$ models trained directly on experimental dpIC$_{50}$ was much lower, at a median $R_{ext}^2 = 0.39$, confirming the large benefit of Profile-QSAR in selectivity prediction.

Surrogate Autoshim. The first step of Ensemble Surrogate Autoshim model building requires selecting a small, diverse set of surrogate structures that perform consistently well in conventional docking, to comprise the S1-serine protease ensemble. An internal collection of ∼1500 S1-serine protease structures from both public and proprietary databases were aligned using the whole protease domain. An initial collection of 22 diverse crystal structures were selected for docking validation. Several factors were considered: (1) diverse sampling of key binding site residues. e.g. residue at the floor of the S_1 pocket, (2) unique loop conformations in the binding site region, e.g. the S_1 loop, (3) unique rotameric states for key residues, e.g. catalytic histidine, S_1 loop residue allowing access to the S_4 pocket, (4) structures with cocrystallized, noncovalent ligands, (5) cocrystallized ligands with few enough rotatable bonds for expedient docking validation studies, and (6) current or future target relevance for small-molecule drug discovery. For some targets, multiple protein structures with cocrystallized ligands in unique binding modes compared to traditional inhibitors, with associated induced-fit protein movements, were included. For several candidates, target diversity was given priority over the docking complexity of the cocrystallized ligand.

To prepare the protein structures for docking, appropriate protonation and tautomeric states were assigned to the ligand−protein complex, followed by restrained minimization to relieve strained geometries. Protein structures initially aligned on the whole protease domain were realigned using a small subset of residues in the binding site region. Residues were chosen where

dx.doi.org/10.1021/ci300059d | J. Chem. Inf. Model. 2012, 52, 2430−2440

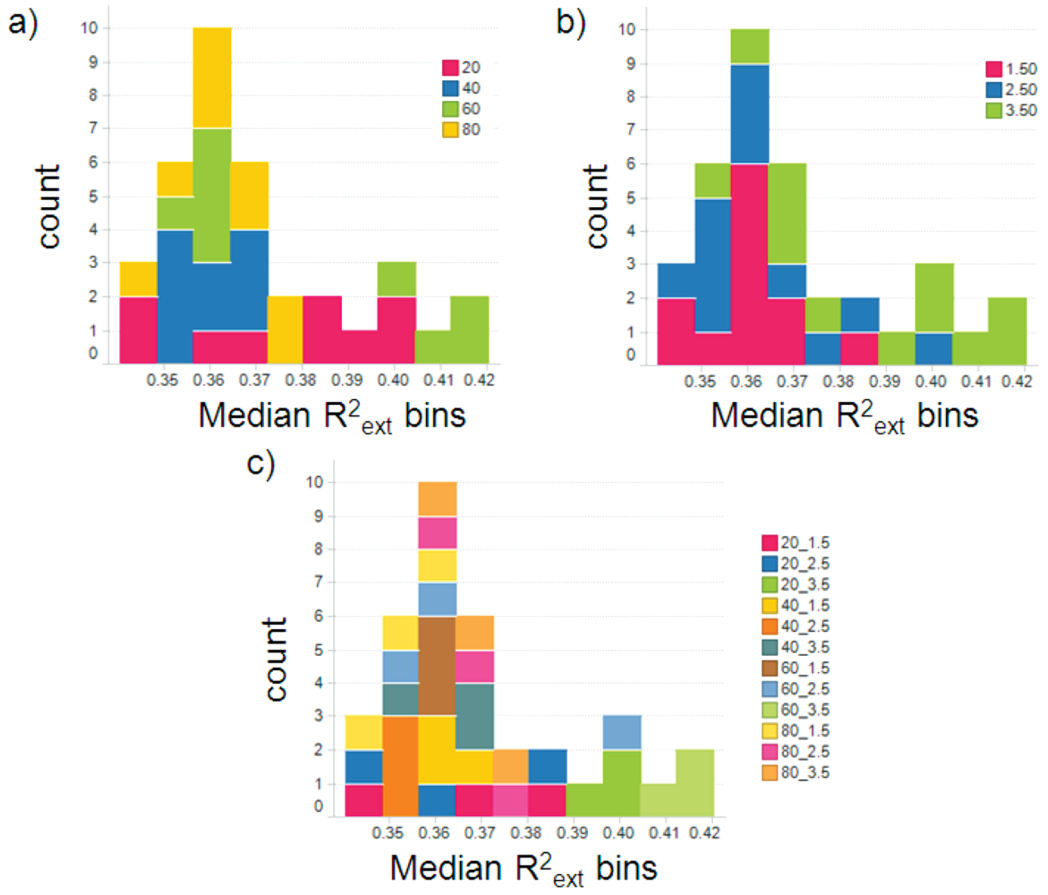


Figure 4. Three histograms show the distribution of model quality, from the 36 runs to optimize the number and sizes of pharmacophore shims, binned by the median R_{ext}^2 correlating predicted and experimental pIC_{50}s for the 24 protease models. The coloring refers to the variations in (a) number of pharmacophore centers, (b) pharmacophore sizes, and (c) combinations of pharmacophore numbers and sizes.

protein−ligand interactions were most frequent, such as S_1', S_1, and S_2 sites, after excluding residues from specific structures whose C_α positions diverged significantly from the remaining structures, e.g. S_1 loop region.

Self-docking pose-prediction studies were conducted using dockit/flo, identical to the previously reported AutoShim workflow. Dockings were done in triplicate, as dockit uses a stochastic method for pose-generation. Modifications to Dockit's default binding-site definitions were required for some structures. Seventeen structures passed the self-docking requirement. For 5 of the 17 structures, the best Flo-scored pose had rmsd ≤2 Å in all (3/3) cases. Three had rmsd ≤2 Å 2/3 times, and 5 had it 1/3 times. Among the remaining 4 structures, 2 had a pose with rmsd ≤2 Å, but it never scored best. The other 2 had ligands with ≥15 rotatable bonds and never generated a pose with rmsd ≤2 Å. Despite this failure, they were kept in the final selection for target diversity considerations. The final selection of 17 structures came from 15 unique proteases.

Coordinates for defining the pharmacophore shim locations were generated by clustering the polar and nonpolar atom coordinates from docked poses at different similarity thresholds (see Methods). To identify a suitable shim configuration, the 8492 compounds from the 24 S1-serine protease assay data sets were docked into the 17 surrogate structures using 36 variations of shim positions and sizes: 4 variations in number of pharmacophore points (20, 40, 60, and 80), by 3 pharmacophore sizes (1.5, 2.5, and 3.5 Å) for the occupancy shims (nonpolar, aromatic and heavy-atom), by 3 replicates to test reproducibility. The 3 histograms from the 36 runs in Figure 4 are binned by the median R_{ext}^2 between predicted and experimental pIC_{50}s for the 24 protease models. Figure 4a, colored by the number of pharmacophore points, shows that the best correlations had 60 shim points (green), followed by 20 shim points (magenta). Figure 4b, colored by nonpolar pharmacophore size, shows that 3.5 Å (green) shims performed best. Figure 4c, colored by combinations of pharmacophore numbers and sizes, shows that the 3 best performing combinations used 60, 3.5 Å shims, with R_{ext}^2 varying from 0.41 to 0.42. This consistency gives confidence that quality of the models is not due to over fitting by testing too many shim definitions. The close second best is 20_3.5 ranking fourth, fifth, and seventh, with R_{ext}^2 varying from 0.39 to 0.40.

Figure 5 plot shows R_{ext}^2 for the 24 S1-serine protease assays for the highest performing shim combination. The performance of Autoshim is shown in blue. For proteases which also had crystal structures in the ensemble, R^2 for the Flo+ docking score is also shown for the test set compounds. The highest R^2 attained through docking was ∼0.3, but the median R_{ext}^2 was 0.02. The median R_{ext}^2 for AutoShim was 0.41, demonstrating a significant boost in predictive performance over conventional docking. In this case, ∼71% of the models passed the threshold of $R_{ext}^2 \geq 0.3$, considered the minimum predictive performance required for prospective use.

Profile-QSAR Evaluation on Cysteine Proteases. The Profile-QSAR methodology was further extended to cysteine

dx.doi.org/10.1021/ci300059d | J. Chem. Inf. Model. 2012, 52, 2430−2440

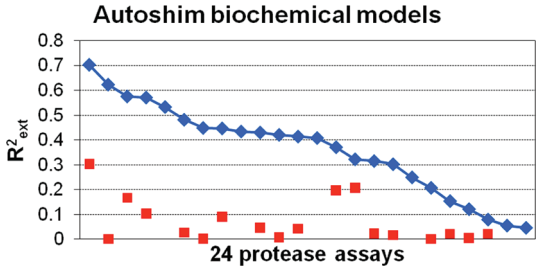


Figure 5. Comparing the performance of Surrogate AutoShim target-customized protease docking to conventional docking with Flo+. AutoShim performance is shown as blue diamonds. R^2 between biochemical pIC_{50} and Flo+ docking scores. The median R_{ext}^2 for AutoShim is 0.41 while that from docking is 0.02.

proteases. Ten cysteine protease assay data sets were selected, only five of which satisfied the usual data criteria for protease autoshim model of ≥250 data points and ≥15 submicromolar inhibitors. These five are referred to as "cysteine large" (C_L), while the five smaller data sets that failed these criteria are referred to as "cysteine small" (C_S). Individual Bayesian models for the 24 S1-serine and 10 cysteine protease assay data sets were trained at up to five activity thresholds as described in Methods and previous publications. Three sets of Profile-QSAR models were trained using various subsets of these Bayesian predicted activities as "chemical descriptors" (see Introduction and Methods): using (a) only the cysteine proteases, (b) only the serine proteases, and (c) the combination of cysteine and serine proteases. Figure 6 shows Profile-QSAR prediction accuracy on the external held-out test sets, measured as R_{ext}^2, for the 34 protease assays. The data sets are divided into three

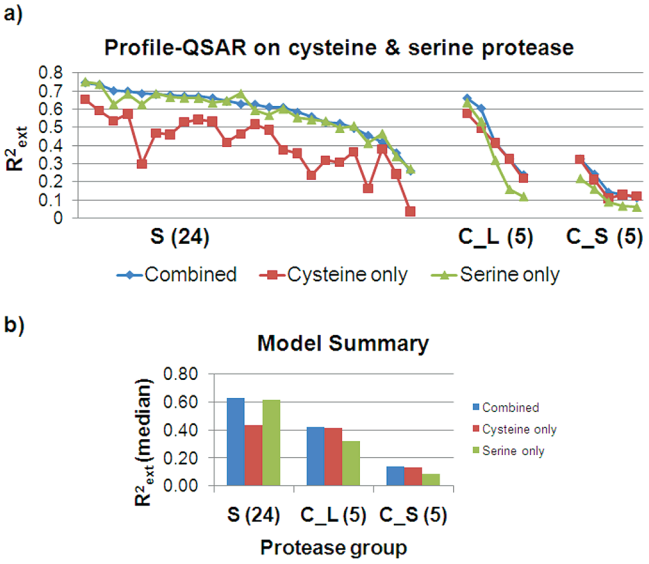


Figure 6. (a) Profile-QSAR prediction accuracy on external held-out test sets, measured as R_{ext}^2 for 24 serine and 10 cysteine proteases. The data sets are divided into three groups on the *X*-axis: S (24) = 24 S1-serine protease data sets, C_L (5) = 5 "large" cysteine protease data sets, and C_S (5) = 5 "small" cysteine protease data sets. The three series, cysteine only (red), serine only (green), and combined (blue), refer to the protease families whose synthetic activities are used as independent variables in the model building. (b) A histogram showing the median R_{ext}^2 for the S, C_L, and C_S data sets using three synthetic activity descriptor combinations (combined serine only, cysteine only).

groups on the *X*-axis: S (24), the 24 S1-serine protease data sets, and C_L and C_S the large and small cysteine protease data sets. The five C_S data sets gave poor performance for all three descriptor combinations, suggesting that the previously established minimum data set criteria were needed to eliminate data sets of insufficient quality. For both the 5 C_L and 24 S data sets, the combination of cysteine and serine protease descriptors gave the best overall performance, with median R_{ext}^2 = 0.63 and 0.42, respectively. The very close next best for C_L and S were their matching descriptor sets, i.e. cysteine only and serine only, respectively, with median R_{ext}^2 = 0.61 and 0.41. The opposing descriptor sets did poorly, with R_{ext}^2 = 0.44 and 0.32 for serine and cysteine, respectively.

DISCUSSION

Protease Profile-QSAR Performance. The 2D Profile-QSAR could simply code inhibitory activity for specific ligand substructural features involved in either nonbonded or covalent interactions. The 3D AutoShim would require special considerations to capture the pharmacophoric and reactivity aspect of functional groups located close to the catalytic residue. Our priority was to identify noncovalent protease inhibitors, so these validation studies removed potential covalent inhibitors. Filtering with a collection of irreversible substructures eliminated only 1.2% of the compounds possibly acting through covalent inhibition, verifying that medicinal chemistry around S1-serine proteases has historically focused on reversible inhibitors.

A full experimental matrix for 24 S1-serine protease assays and 8492 compounds would contain 203 808 experimental IC_{50}s. The protease assay data sets included 20 610 IC_{50}s, only ∼10% full. While the sparseness of the protease activity matrix was similar to that found for kinases, the overall numbers were much lower. The kinase activity matrix included 1.5 million IC_{50}s for 130 000 compounds across 115 kinase assays. The limited data meant that the criteria for selecting data sets for modeling in kinases, i.e. ≥600 data points and ≥15 submicromolar compounds, would be too strict for proteases. Lower thresholds of ≥250 IC_{50}s and ≥15 submicromolar compounds resulted in 24 S1-serine protease assay data sets for modeling.

As described in the Results section, protease Profile-QSAR yielded a median R_{ext}^2 = 0.60 for 24 S1-serine protease data sets ranging in size from 273 to 3366 IC_{50}s. The performance of Profile-QSAR on 115 kinase assays ranging is size from ∼700 to ∼58 000 data points gave a comparable median R_{ext}^2 = 0.59, demonstrating that accurate Profile-QSAR models can be trained on limited activity data. The results of the Profile-QSAR selectivity modeling were similar to what has been observed previously for kinases.[1] The Profile-QSAR "delta models" performed slightly better than the "difference models", but both models performed significantly better than the Bayesian QSAR delta models.

Only five large cysteine protease data sets met the minimum data set criteria for model building, Models with profiles from five small cysteine protease data sets were also attempted, but the poor results for these data sets confirmed the previously set criteria for minimum data set requirements. Adding the Bayesian models from the 24 serine proteases to the five from the large cysteine proteases slightly improved the two best cysteine protease models; although with so few examples it is hard to generalize. Even so, with median R_{ext}^2 = 0.42, the performance is much weaker than a median R_{ext}^2 = 0.62 for the

S1-serine proteases. However, since the combined descriptors cover more chemical space than either set alone, and there appears to be no penalty, the former was chosen for future cysteine protease Profile-QSAR modeling.

Around 80% of the S1-serine proteases being evaluated for model building are extracellular targets. Therefore, at present, sufficient data does not exist for cellular Profile-QSAR modeling of S1-serine protease inhibitors.

Protease AutoShim Performance. The Autoshim modeling used a 17 structure S1-serine protease ensemble selected from a larger collection of ~1500 structures. Variations in the number of pharmacophoric shims and the size of occupancy shims were evaluated. The combination of 60 shims of 3.5 Å was best, with median R_{ext}^2 ranging from 0.40 to 0.41, but 20 features of 3.5 Å was a close second with median R_{ext}^2 ranging from 0.39 to 0.41. The latter is very similar to the kinase implementation.[2]

While protease Profile-QSAR performance was comparable to kinases, for AutoShim, the median R_{ext}^2 = 0.41 for 24 S1-serine protease data sets is lower than typical for kinases. There are many differences between kinase docking and protease docking, but limited training data and lack of potent compounds alone might account for much of the decline. Previous kinase AutoShim[2] and Profile-QSAR[1] publications noted that the dynamic range, measured as σpIC_{50}, and data set size to a lesser extent, correlated model predictive performance measured as R_{ext}^2. To test whether the more limit protease training data accounted for the quality difference, σpIC_{50} and R_{ext}^2 from 71 kinase models with a median R_{ext}^2 = 0.45 was compared to that of the 24 S1-serine protease models. The kinase data sets all passed the criteria of ≥600 data points and ≥15 submicromolar compounds with the number of data points ranging from 684 to 58 643 and number of submicromolar compounds ranging from 22 to 10 201. The protease data sets passed the lesser criteria of ≥250 data points and ≥15 submicromolar compounds with number of data points ranging from 372 to 3366 and number of submicromolar compounds ranging from 24 to 820. Figure 7 plots R_{ext}^2 on the

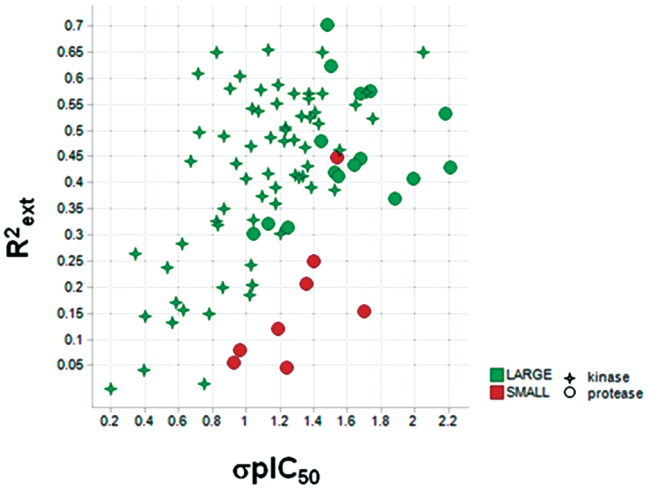


Figure 7. Plot comparing the relationship between predictive power (R_{ext}^2) dynamic range (σpIC_{50}), and data set size for 51 kinase (star) vs 24 S1-serine protease (circle) models. Small data sets meeting the criteria of ≤900 data points and ≤100 submicromolar compounds are shown in red, and large data sets not meeting the criteria are shown in green.

Y-axis and σpIC_{50} on the *X*-axis for the 71 kinase (shown as stars) and 24 protease models (shown as circles). Small data sets with ≤900 data points and ≤100 submicromolar compounds are in red, while the remaining large data sets are in green. The large protease data sets are mixed among the large kinase data sets, although the majority lies toward the lower edge. The small data sets are all proteases, and all but one fall below the large data sets, both protease and kinase. Increasing the sizes of these data sets and adding more submicromolar inhibitors should significantly improve their performance to match the large protease data sets. It is also appears AutoShim models require larger data sets than Profile-QSAR.

Cysteine Protease Models. Only five large cysteine protease data sets met the minimum data set criteria for model building, roughly one-fifth of the 24 S1-serine protease data sets suitable model building. Five small cysteine protease data sets were also evaluated, but the poor results for these data sets validated the previously set criteria for minimum data set requirements.

The best performance of Profile-QSAR on the large cysteine protease assay data sets, with median R_{ext}^2 = 0.42, obtained combining serine and cysteine protease synthetic IC_{50} predictions. This was modest compared to the median R_{ext}^2 = 0.62 for the S1-serine proteases. The performance of the combined bayesian predictions from the S1-serine and cysteine proteases was comparable to that from using only the intrafamily S1-serine and cysteine protease synthetic IC_{50}s. Since using the combined synthetic IC_{50}s is no worse than the individual predictions and provides a wider chemical space coverage than the individual bayesian descriptor sets, the former would be used for future Profile-QSAR modeling around these target-families.

The partial-least-squares process generates a regression equation for the activity as a linear combination of orthogonal "latent variables", some positive and some negative. The latent variables are themselves linear combinations of the original Bayesian activities, which best correlate with both the descriptor matrix and the target IC_{50}s. In Profile-QSAR, the original variables were themselves modeled activities based on substructural fingerprints. This complex process makes it difficult to identify the structural features responsible for activity. However, basis set kinases which are strongly correlated or anticorrelated with the target kinase should contribute significantly toward that model. This is seen in Figure 6, where serine protease activity is better predicted by a basis set of serine protease models than a basis set of cysteine protease models and vice versa.

Prospective Project Application. The first protease application was for a cysteine protease, "C1", which is an important target in the oncology and inflammation area and had significant medicinal chemistry invested in two scaffolds. Unfortunately, both of these scaffolds had flat SAR profiles and ADME liabilities, which motivated the team to look for new chemical matter. Bayesian QSAR built on only C1 data gave a modest R_{ext}^2 = 0.23. However, a Profile-QSAR model combing the C1 bayesian and S1-serine protease bayesian descriptors gave a much improved model with R_{ext}^2 = 0.41, demonstrating that the S1-serine synthetic IC_{50}s improved the predictive quality of the model.

Activity was predicted for 1.5 million compounds from the corporate archive Compounds with a predicted $pIC_{50} \geq 5$ were considered further (Figure 8a). Compounds with Tanimoto

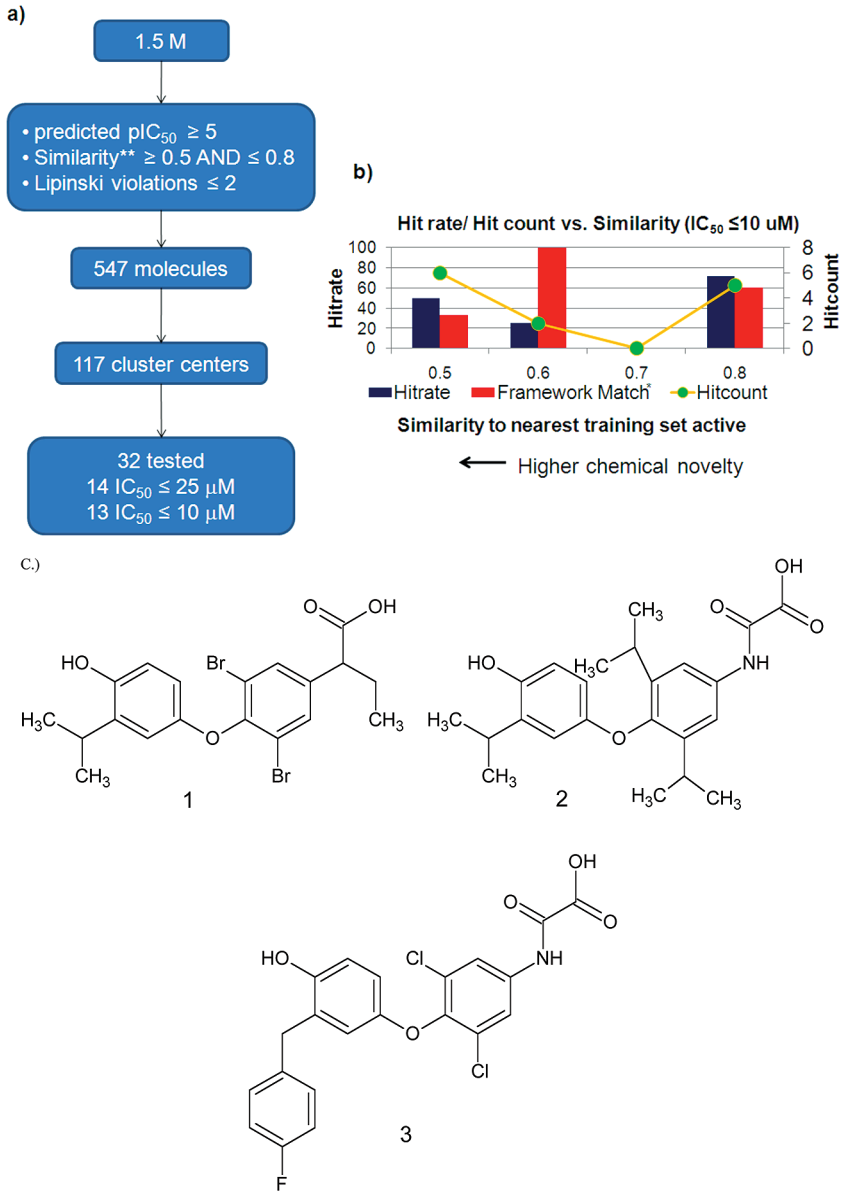


Figure 8. (a) Workflow used for virtual screening of cysteine protease "C1". (**) Tanimoto similarity between each putative hit and the closest training set active with experimental $pIC_{50} \geq 5$. (b) "Chemical novelty histogram" with the bin of Tanimoto similarity to the nearest training set active ($pIC_{50} \geq 5$) on the *X*-axis, hit-rates on the primary *Y*-axis, and hit count on the secondary *Y*-axis. The blue bars show hit-rates, while the red bars show "framework match", i.e. (*) percentage of hits in each bin which share a Bemis and Murcko framework match with a framework found among the training set actives. The green circles show the hit counts for each similarity bin. (c) Structures of three compounds selected through the virtual screening and evaluated against C1.

similarity ≥0.9 to a member of the training set would give high hit-rates, but the goal was novelty; so, these compounds were discarded. Conversely, with such limited coverage of protease chemical space, extreme extrapolation would likely result in low hit-rates. Therefore, a similarity filter selected compounds with Tanimoto similarity of ≤0.8 and ≥0.5 to the closest active training set member. Removing molecules with >2 Lipinski violations left 547 compounds, which were classified into 117 clusters. Thirty-two cluster centers were ordered and tested, of which 13 were active at $IC_{50} \leq 10$ μM. This 41% hit-rate was only slightly lower than the median hit-rate of 50% for 8 kinase applications.[1] Figure 8b shows that 8 of the 13 hits are in the most novel 0.5 and 0.6 similarity bins. Four had unique scaffolds unrepresented in the training set. Three of the novel scaffolds are being followed up by testing additional analogues. Figure 8c shows three others of the 13 active compounds. Compound 1 had $IC_{50} > 3.2$ μM when retested in a lower concentration rundown. Compounds 2 and 3 had $IC_{50} = 2.5$ and 0.35 μM, respectively.

An effort to prioritize C1 scaffolds by identifying chemotypes sharing similar SAR profiles in target cysteine protease C1 and 13 off-target S1-serine proteases required IC_{50}s for all compounds against the 14 proteases. The compound by protease activity matrix was only 7.5% full, of which 7.2% came from the experimental C1 data, and only 0.3% came from the other 13 proteases. Profile-QSAR predictions filled in the remaining 92.5% of pIC_{50}s. The 1959 compounds in the C1 data set was clustered based on both chemical structure and cross-reactivity profile, measured as correlation squared (R^2) between C1 experimental pIC_{50} and mixed experimental/

dx.doi.org/10.1021/ci300059d | J. Chem. Inf. Model. 2012, 52, 2430−2440

predicted pIC$_{50}$s of the 13 off-target serine proteases. Figure 9 shows R^2 on the *Y*-axis, and a hierarchical classification on the

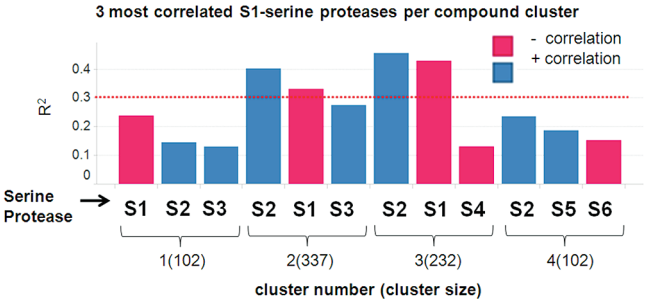


Figure 9. Bar graph showing R^2 on the *Y*-axis. On the *X*-axis, a hierarchical clustering is divided first by cluster number and cluster size, and then by individual serine proteases (S1−S6). The three proteases with the highest correlations for each of the four structural clusters are shown. Within each structural cluster, the three highly correlated proteases are arranged in decreasing order of R^2. Positive correlations are shown in blue while negative correlations are shown in red. The red dotted line represents a threshold of R^2 = 0.3 which is considered the lowest threshold for significant correlation.

X-axis, for the four largest compound clusters. For each compound cluster, the three S1-serine proteases with the highest correlation to C1 are plotted, arranged in descending order. Positive correlations are shown in blue while negative correlations are shown in red. Serine protease "S2" has the highest positive correlation in all cases, portending selectivity difficulties. S1 has a large negative correlation in three of four cases, so selectivity should not be problematic. The correlations are generally low, but this is not surprising with so many predicted activities, although correlations below 0.3 are suspect even with predicted activities. Especially if chemotypes from clusters 2 or 3 are studied, careful experimental profiling in S2 would be advised. Conversely, scaffolds established in the S2 program might be repurposed through medicinal chemistry toward C1.

CONCLUSION

Two protein-family iterative virtual screening methods, 2D Profile-QSAR and 3D Autoshim, together form an IMTS in silico alternative to HTS. The methodologies have been successfully applied to over two dozen active kinase projects in a variety of roles: target tailored virtual screening, HTS triaging, isoform and pan-kinome selectivity profiling, target similarity analysis, and understanding disconnects between cellular and biochemical end-points. Kinases are particularly amenable to the protein-family approach, since they share structural similarity of the ATP binding-site with corresponding cross-reactivities. The protein-family models repurpose the huge wealth of historical activity and structural data for applications to new targets within the protein-family. Proteases, particularly S1-serine proteases, provide some of the same advantages and were chosen for extension of the IMTS methodologies.

Extension to S1-serine proteases was validated on 24 IC$_{50}$ data sets, from 16 unique enzymes, which were prefiltered to remove potential covalent inhibitors. Protease activity data were far fewer than kinases, so lower data set minimum criteria were employed. 2D Profile-QSAR models on these 24 assay data sets gave excellent performance, with median R_{ext}^2 = 0.60, similar to that observed for kinases. Performance greatly exceeded the corresponding Bayesian-QSAR models, with median R_{ext}^2 = 0.40. Two kinds of Profile-QSAR selectivity models, "delta models" and "difference models", gave excellent results for 60 pairs of proteases, with median R_{ext}^2 = 0.64 and 0.61, respectively. These were significant improvements over rational Bayesian-QSAR models. Since over 80% of the current S1-serine protease targets are extracellular, cellular activity modeling was not undertaken.

A Profile-QSAR extension was also attempted for cysteine proteases. Although ten cysteine protease assay data sets were available, only five data sets met the minimum requirements. Results from this limited study showed that overall performance for the cysteine protease models were significantly lower, but still useful, with a median R_{ext}^2 = 0.42.

A 17 structure S1-serine protease surrogate ensemble for 3D AutoShim modeling was generated from a starting collection of ∼1500 X-ray structures. Optimizing the pharmacophoric shims yielded a median R_{ext}^2 = 0.41. This performance is lower than observed for kinases. Higher structural diversity of protease binding sites and smaller protease IC$_{50}$ data sets with fewer potent compounds explains to the lower performance.

Serine proteases share similar features with cysteine proteases, and the S1-serine protease Profile-QSAR was successfully employed in iterative virtual screening for the "C1" cysteine protease project. Subsequently, a general cysteine protease Profile-QSAR model was developed using combined synthetic IC$_{50}$s from both S1-serine proteases and cysteine proteases. Predictive performance was very slightly higher than using the cysteine protease synthetic IC$_{50}$s alone. Virtual screening with the model led to ordering and screening 32 compounds, which provided 13 hits at IC$_{50}$ ≤ 10 μM, a 41% hit-rate. Three new scaffolds were identified, which are being followed-up with testing of additional analogues. A cross-reactivity analysis using mainly Profile-QSAR predicted activities for 13 S1-serine proteases revealed that, at $R^2 \geq 0.3$, one serine protease showed consistent positive correlation and one showed consistent negative correlation, for the two C1 chemotypes with the largest number of synthesized analogues.

Overall, as with the kinases, extension of Profile-QSAR and Surrogate AutoShim to serine and cysteine proteases showed unprecedented virtual screening accuracy. One fruitful application has been achieved, with the promise of many more to come.

ASSOCIATED CONTENT

Supporting Information

Identity of selected protease data sets used in the validation. Identity of the publicly available crystal structures used in the AutoShim model building. Cartesian coordinates of a 60 point pharmacophore used in the validation study. Alignment of the publicly available crystal structures used in the protease ensemble and a .sea file for the 60 point pharmacophore definitions as defined for the Magnet program. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: prasenjit.mukherjee@novartis.com.

Present Address

†Structural Research, Department of Medicinal Chemistry, Boehringer-Ingelheim, 900 Ridgebury Road, Ridgefield, CT 06877, USA.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

P.M. would like to thank the NIBR Education office for postdoctoral funding.

■ REFERENCES

(1) Martin, E.; Mukherjee, P.; Sullivan, D.; Jansen, J. 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.

(2) 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.

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

(4) 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* **2010**, *468*, 973−977.

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

(6) Shi, H.; Kong, X.; Ribas, A.; Lo, R. S. Combinatorial Treatments That Overcome PDGFR.beta.-Driven Resistance of Melanoma Cells to V600EB-RAF Inhibition. *Cancer Res.* **2011**, *71*, 5067−5074.

(7) Liao, J. J.-L. Molecular Recognition of Protein Kinase Binding Pockets for Design of Potent and Selective Kinase Inhibitors. *J. Med. Chem.* **2007**, *50*, 409−424.

(8) Rawlings, N. D.; Barrett, A. J. Evolutionary families of peptidases. *Biochem. J.* **1993**, *290*, 205−218.

(9) Rawlings, N. D.; Tolle, D. P.; Barrett, A. J. MEROPS: the peptidase database. *Nucleic Acids Res.* **2004**, *32*, D160−D164.

(10) Puente, X. S.; Sanchez, L. M.; Overall, C. M.; Lopez-Otin, C. Human and mouse proteases: A comparative genomic approach. *Nat. Rev. Genet.* **2003**, *4*, 544−558.

(11) Drag, M.; Salvesen, G. S. Emerging principles in protease-based drug discovery. *Nat. Rev. Drug Discovery* **2010**, *9*, 690−701.

(12) Turk, B. Targeting proteases: successes, failures and future prospects. *Nat. Rev. Drug Discovery* **2006**, *5*, 785−799.

(13) Abbenante, G.; Fairlie, D. P. Protease inhibitors in the clinic. *Med. Chem.* **2005**, *1*, 71−104.

(14) Acharya, K. R.; Sturrock, E. D.; Rirodan, J. F.; Ehlers, M. R. W. ACE revisited: A new target for structure-based drug design. *Nat. Rev. Drug Discovery* **2003**, *2*, 891−902.

(15) De Clercq, E. Antiviral drugs in current clinical use. *J. Clin. Virol.* **2004**, *30*, 115−133.

(16) Ruef, J.; Katus, H. A. New antithrombotic drugs on the horizon. *Expert Opin. Investig. Drugs* **2003**, *12*, 781−797.

(17) Gustafsson, D.; Bylund, R.; Antonsson, T.; Nilsson, I.; Nystroem, J. E.; Eriksson, U.; Bredberg, U.; Teger-Nilsson, A. C. Case history: A new oral anticoagulant: the 50-year challenge. *Nat. Rev. Drug Discovery* **2004**, *3*, 649−659.

(18) Overall, C. M.; Lopez-Otin, C. Strategies for MMP inhibition in cancer: innovations for the post-trial era. *Nat. Rev. Cancer* **2002**, *2*, 657−672.

(19) Powers, J. C.; Asgian, J. L.; Ekici, O. D.; James, K. E. Irreversible inhibitors of serine, cysteine, and threonine proteases. *Chem. Rev.* **2002**, *102*, 4639−4750.

(20) Hedstrom, L. Serine Protease Mechanism and Specificity. *Chem. Rev.* **2002**, *102*, 4501−4523.

(21) http://en.wikipedia.org/wiki/Naive_Bayes_classifier (accessed 3/1/2011).

(22) *Pipeline Pilot v 8.0 data modeling user guide*; 2011.

(23) Bemis, G. W.; Murcko, M. A. Properties of Known Drugs. 2. Side Chains. *J. Med. Chem.* **1999**, *42*, 5095−5099.

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