Surrogate AutoShim: Predocking into a Universal Ensemble Kinase Receptor for Three Dimensional Activity Prediction, Very Quickly, without a Crystal Structure

Eric J. Martin* and David C. Sullivan

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

Received December 7, 2007

"Ensemble surrogate AutoShim" is a kinase specific extension of the AutoShim docking method that solves the three traditional limitations of conventional docking: (1) it gives good correlations with affinity, (2) does not require a target protein structure, and (3) for a preprocessed company archive of 1.5 million compounds, is as fast as traditional 2D QSAR. It does require several hundred experimental IC_{50} values for each new target. Original AutoShim adds pharmacophore "shims" to a crystal structure binding site. An iterative partial least squares (PLS) procedure selects the best pose, while adjusting the shim weights to reproduce IC_{50} data. *Surrogate* AutoShim adjusts shims in one crystal structure to reproduce IC_{50} data for a different kinase target. *Ensemble* surrogate AutoShim uses 16 structurally diverse kinase crystal structures as a "universal ensemble kinase receptor", suitable for any kinase target. The 1.5 million member Novartis screening collection has been predocked into the shimmed ensemble, so new kinase models can be built, and the entire corporate archive virtually screened, in hours rather than weeks. A kinase-biased set of 10 000 compounds, that samples the entire corporate archive, has been designed for lead discovery by iterative kinase screening.

INTRODUCTION AND OVERVIEW

Limitations of Docking with All-Purpose Scoring Functions for Predicting Activity. Although docking is one of the most commonly used methods in computational chemistry, all-purpose scoring functions generally cannot predict IC_{50}. While widely recognized for many years, the extent of the challenge was highlighted in the recent publication of a huge docking survey at GlaxoSmithKline.[1] This paper, with 15 authors, studied 8 protein targets, using 10 docking programs, each with several scoring functions. While the study did find useful pose predictions, and modest enrichments in predicting yes/no activity, the following quote summarizes the disappointing results regarding affinity prediction: "there is no statistically significant correlation between measured affinity and any of the [37] scoring functions [from 10 docking programs] evaluated across all eight protein targets examined (Table 7). An extremely modest positive correlation was observed for Chk1 [$R^2 = 0.32$]." Ensemble surrogate AutoShim aims to improve on conventional docking, even in the absence of a crystal structure. Routinely exceeding this best value of $R^2 = 0.32$ from the nearly 300 attempts in the GSK study will be taken as our criterion for success.

AutoShim. If a crystal structure is available, original "AutoShim" creates a target-specific scoring functions to address the limited affinity prediction of general-purpose scoring functions.[2] AutoShim is an automated extension of the Magnet postdocking analysis program, an interactive expert system which defines and evaluates specific interactions between a bound ligand and a protein active site.[3] Analogous to an NMR magnet, AutoShim's automated c2sea program adds Magnet pharmacophore "shims" to the protein active site, and Magnet evaluates how each docked pose interacts with the shims. This can be a series of simple single-point pharmacophore shims, or a series of complex multipoint shims generated using recursive partitioning (RP). Partial least squares (PLS) regression weights the shims, along with the general-purpose docking score and other terms, to create a new adjusted scoring function that optimally reproduces experimental activity data. A robust iterative procedure is employed to combine pose selection with model training. It converges in only 2–4 cycles, thereby avoiding overtraining. It gives comparable models irrespective of starting poses or training sets. In six of the eight published AutoShim examples using diverse compound sets, the resulting models exceeded $R^2 = 0.32$, the best correlation from the nearly 300 attempts in the GSK study above. Pure single-point pharmacophore models are slightly less predictive than corresponding multifeature shim models, but the shim weights can be interpreted to identify desirable features for tight-binding ligands.

Surrogate versus Homology Docking. Conventional docking requires a protein structural model, ideally generated from high-resolution crystallographic data of the protein bound to a ligand similar to those to be docked. In cases where a suitable experimental structure for a particular target is not available, one can dock instead into the closest homologue with a solved structure ("surrogate docking"). Alternatively, this homologue can serve as a template for building a 3D-protein model from the target sequence ("homology docking"). While the comparative modeling step can add value to the docking calculation,[4] in many cases docking directly into the template performs as well or better,

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

10.1021/ci700455u CCC: $40.75
Published on Web 04/02/2008

while avoiding the possibility of adding artifacts that can significantly reduce accuracy.[5]

As in AutoShim, where shims were added to compensate for errors in a general-purpose scoring function, surrogate AutoShim adds shims to the docking results from an available crystal structure to create a new scoring function that will reproduce activity data for a different but related protein that lacks an experimental structure, i.e. it shims the surrogate active site to behave like the target of interest. For the special case of the protein kinases, which all have closely related active sites, there are many suitable surrogates for any of the 518 human family members.

Protein Flexibility. Compounding the structural uncertainty surrounding kinases without crystal structures (or even those with), protein structures are flexible. Protein flexibility poses a particular issue for kinase active sites, which sit at a "hinge" between two domains, and are lined by the activation loop, which shifts dramatically upon phosphorylation. Ligand binding significantly influences active site structure,[6] and environmental factors unrelated to biological activity, such as crystal packing, may further perturb structure.[7] Diller comments that interkinase structural variation typically is not significantly more than intrakinase structural variation.[8] The relatively low importance of high sequence-similarity in determining active site structure explains the poor correspondence between binding site sequence similarity and actives-enrichment in homology docking.[9]

Ensemble AutoShim and Ensemble Surrogate AutoShim. Original AutoShim deals with protein flexibility when several crystal structures with diverse binding site conformations are available. The ligands are docked into each structure separately, and the structures are superimposed before the shims are added. The protein structure is then simply treated as part of the "pose" during the iterative pose-selection/model parametrization process.

Ensemble surrogate AutoShim extends this approach to cases where a single structure, or even no structure, is available. A "universal ensemble kinase receptor" (UEKR) was made by superimposing 16 of the most diverse crystal structures available, comprised from 14 different kinases. This model is intended to cover the full range of conformations available to any kinase. Compounds are docked into each of the 16 structures, and a new ensemble surrogate AutoShim scoring function is "shimmed" to reproduce experimental activity data for any new kinase, just as if they were various conformations of the target kinase itself. This can be done even without a crystal structure of the actual target. If a crystal structure does exist, adding the ensemble to it yields improved models, presumably by better accounting for flexibility.

Predocking for High Throughput. All of the slow 3D computations in ensemble surrogate AutoShim are performed at the outset, before the experimental IC_{50} data are introduced: the docking, conventional scoring, and the evaluation of each pose for pharmacophore atom counts and H-bond distances with Magnet. At that point, the 3D docking coordinates are no longer needed, having been captured in a vector of descriptors measuring the interactions of each pose with the 120 Magnet features. The remaining target-specific steps (combining features into target specific shims, scoring-function parametrization by PLS with iterative pose selection, model validation on a held-out test set, and predictions for unknowns) are performed on these Magnet feature vectors, and are comparable to conventional QSAR with 2D descriptors. Because the UEKR always uses the same 16 diverse kinase crystal structures, the slow 3D docking and Magnet steps need only be done once for any collection of compounds, and the Magnet feature vectors can then be used for every new kinase target. We have predocked and "Magnetized" the 1.5 million member Novartis screening collection in the 16 crystal structures of the UEKR. This took months on a Beowulf cluster, but now ensemble surrogate AutoShim docking for each new kinase target is as fast as conventional QSAR.

An iterative screening procedure has been developed on this technology to take advantage of the new "cherry picking" robots. A kinase-biased set of 10 000 compounds has been designed to sample the entire corporate archive. For any new kinase target, an initial screen is performed with these 10 000 compounds. An ensemble surrogate AutoShim model is parametrized from these experimental affinities using the predocked poses. If the model is predictive on a held-out test set, it is used for a virtual screen of the entire 1.5 million predocked compounds. Because no docking is involved, this takes days instead of weeks. Several thousand predicted actives are retrieved and tested with the newly available cherry picking robots to find new hits. These additional results can then be added to the original training set to refine the surrogate AutoShim model and find additional hits.

METHODS

Docking. Prior to docking, 16 crystal structures were placed in a common coordinate system by superimposing active sites using the backbone atoms of the hinge residues and the aspartate of the DFG loop. For each target, Dockit[10] generated an initial set of 150 poses, of which the 100 top scoring were retained. These poses were energy minimized in the active site with Flo+.[11] Following docking and minimization, Magnet removed poses that did not make at least one hydrogen bond to the hinge backbone.[3] Poses with severe steric clashes following minimization were likewise removed.

Compound Selection. Here, 71 280 compounds with either in-house or outsourced dose–response inhibition (IC_{50}) data against any of 68 kinases were docked into the 16 crystal structures of the UEKR (see below). Following docking, minimization and filtering, 71 211 compounds emerged from at least 1 of the 16 docking models, with 87 average total poses divided across the 16 docking models. These data sets contained both medicinal chemistry series and diverse purchased compounds. They are overall diverse, similar to the sets characterized more fully in the companion article on AutoShim.[2]

Redundant Pose Filtering. For single-target (i.e., non-ensemble) surrogate AutoShim modeling, a subset of structurally diverse yet energetically favored conformations was selected by ranking the docking model's generated pose set by Flo+ score for each compound. Poses where no corresponding atom pairs differed by 3 Å from the top-ranked conformation were eliminated as redundant. This operation was repeated with the next-best remaining pose, and so on down the ranked list. A maximum of nine top-Flo+ scoring diverse poses were carried forward to surrogate AutoShim.

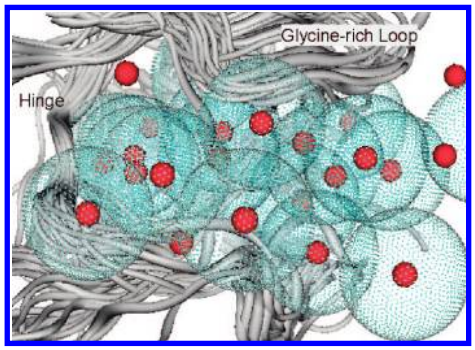

Figure 1. Sixteen target overlay with shim positions. The 16 UEKR backbone is shown with the "atom-count" spheres (blue dots), which have a radius of 3.5 Å, and the hydrogen-bond distance shim centers (small red spheres).

For UEKR AutoShim modeling, each compound's poses across all 16 docking models were pooled, ranked by Flo+ score, and the nine diverse/high-Flo+ scoring compounds selected by the same procedure as for single-target surrogate AutoShim.

Automated Generation of Magnet Interaction Features: c2sea and Extraction for Each Pose. The combined active site volume across the diverse kinase ensemble is much larger than any individual target. Compared to the single targets in our accompanying manuscript,[2] the UEKR would require larger radii spheres about shim positions, or more shim points. The small binary features used in original AutoShim were replaced with larger, more information-rich interaction features that benefit from subsequent recursive partitioning's (RP) ability to find optimal split points in continuous variables. Rather than recording the mere absence/presence of aromatic or hydrophobic atom-types, occurrences were counted in fixed radius spheres of 3.5 Å for three atom-types: aromatic, hydrophobic, and nonhydrogen. For hydrogen bond donors and acceptors (HBD and HBA), position is generally more important than their count in a region of the active site. Therefore, HBD and HBA features record the *distance* from each shim position to the nearest ligand HBD or HBA atom. Ligands lacking any HBD or HBA have this descriptor set arbitrarily large for every shim point. While the raw distance value might trend with activity in a complicated or nonlinear fashion, we rely on subsequent recursive partitioning to choose an optimal count or distance. Feature positions were generated by clustering atom positions from a sampling of ligands docked into the 16 superimposed active sites. The feature points are thus identical for all 16 aligned protein structures. There are 30 HBD or HBA positions and 20 atom-count feature positions. Figure 1 locates feature positions relative to the 16-member active site ensemble. The C program c2sea automatically defines these 120 total interaction features in the SEA language that is understood by the program Magnet.[3] Magnet evaluates the atom counts and H-bond distances for each pose. "Magnet features" and shims are easily confused. "Features" refer to the 120 atom counts and H-bond distances identified by Magnet for each pose. These are independent of the IC_{50} training data. RP of the features against the IC_{50} data picks optimal count and distance thresholds, and feature combinations, to yield target-specific, multipoint, yes/no, pharmacophore shims, for use as PLS descriptors in either ensemble or single-target surrogate AutoShim models (see below).

Generating RP Optimized Multipoint Binary Shims from Magnet Interaction Features. Ensemble and single-target surrogate AutoShim models in this study use extensions of the PLS-on-RP optimized multipoint pharmacophore shims developed in the accompanying AutoShim manuscript.[2] Whereas original AutoShim used RP to build multipoint pharmacophore shims from simple presence/absence Magnet interaction features, Surrogate AutoShim uses the 120 Magnet count or distance interaction features, Flo+ score, and ligand nonhydrogen atom count, as continuous variables for building six recursive partitioning trees, each tree applying a different activity threshold for classifying actives and inactives. These six trees, each with a maximum of six branches, generate sets of multipoint pharmacophores of up to seven Magnet interaction features each, with optimal choices of atom count or H-bond distance thresholds for each shim, which best distinguish active compounds. Note that each feature can occur in many shims. The other significant parameters controlling tree-branch depth in the rpart function, from the rpart package of the R statistical environment, include the minimum node population (set to six) and complexity parameter (0.01). The individual optimized interaction features comprising the multipoint pharmacophores are gathered from a seventh tree, with a fixed pIC_{50} activity threshold of 5.8, to use as optimized single-point atom count or H-bond distance pharmacophores. The atom counts and H-bond distances previously extracted by Magnet for each pose are then compared to these RP optimal thresholds to yield a yes/no descriptor value for each single or multipoint pharmacophore shim. At this point, each pose is now described by a binary shim-*fingerprint* of several hundred bits indicating which multipoint and single-point optimal count and distance pharmacophore shims it satisfies. Note that the 120 atom count and H-bond distance *features* originally extracted by Magnet from the dockings depend only on the ligand pose and protein structure into which the ligand was docked, but not on the kinase target for which a customized surrogate scoring function is being developed. However, each pose's *shim set* recording which single- and multipoint shims are to be matched (including matching the RP optimized count and distance thresholds) by those magnet features, has been trained by RP on IC_{50} data for each new kinase target, and is thus specific for that kinase of interest. Thus, each pose has a different shim-*fingerprint* for each new target, indicating which members of the shim set were matched by its feature set.

PLS Regression. Using the IC_{50} training data, the binary shim-fingerprint descriptors are combined with the continuous raw Flo+ score, atom counts, and contact terms, for PLS regression, using the mvr function of R's pls package, to create a target cutomized AutoShim scoring function. For ensemble surrogate AutoShim models, the compound's 16 best Flo+ scores, 1 for each structure, are also included as descriptors in the PLS regression. Unsuccessful dockings are filled with –9, which is approximately the lowest Flo+ score observed.

RP, which provides the multifeature shims used as binary PLS descriptors, is performed outside the PLS cross-validation step. This could result in overtraining, with PLS cross-validation selecting more latent variables than optimal. As an ad hoc measure to limit PLS overtraining, the

correlation calculated during cross-validation, Q^2, is penalized by 0.002 multiplied by the number of latent variables.

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

Iterative Interleaved Model Training and Pose Selection. As in original AutoShim,[2] the multiple pose problem is solved by applying iterative pose-selection/PLS regression cycles. Initially, PLS regression is applied to the best Flo+ scoring pose to train an interim surrogate AutoShim scoring function. In subsequent iterations, the best scoring pose from the previous AutoShim PLS model trains the next PLS model. All AutoShim models are trained for 10 iterations, although they usually converge in 2–4. For the superimposed ensemble models, all poses from all 16 protein structures are simply combined into one pool of poses, where a "pose" now includes a protein structure as well as a conformation and orientation of the ligand.

Simple Surrogates. AutoShim models were trained for eight kinase targets. In addition to training AutoShim models with pIC_{50} data from the native kinase target, each activity data set also trained AutoShim models using docked poses from the other seven targets, generating one "native" and seven "surrogate" AutoShim models per activity data set.

Ensemble of 16 Active Sites. The ensemble of eight kinases that built simple surrogate AutoShim models was expanded to 14 by selecting crystal structures that cover the range of kinase active site geometries. Two kinases contributed two structures apiece, yielding a total of 16 active sites. This superimposed ensemble serves as a universal ensemble kinase receptor (UEKR) that attempts to capture all the available conformations of any single target. Dockings against all 16 structures are pooled. At this point, shims are added, poses evaluated by Magnet,[3] and *ensemble* surrogate AutoShim models are trained on activity data from each new single kinase target, just as if they had been docked into an ensemble of that target's crystal structures.

Profile Boost. Having affinity prediction models for a collection of kinases enables "ensemble models"[12] that leverage cross-kinome affinity correlations. "Profile boosted AutoShim" employs PLS to model a single kinase's pIC_{50} data as a linear combination of AutoShim predictions from all previous kinase models. Prediction accuracy is evaluated using the same 25% test set of compounds, withheld from both AutoShim training and profile boosting.

One-Time Big Predock. The docking steps using the UEKR are the same irrespective of the kinase target of interest: the docking into 16 protein structures, ligand geometry optimization, conventional scoring, and the feature extraction, via Magnet, of atom counts and H-bond distances for each pose. Once this has been done, these same magnet features are used for making RP optimized shims, corresponding binary shim-fingerprints for each pose, and subsequent PLS scoring functions and activity predictions for any ensemble surrogate docking model, trained, on any new target activity data. Toward that end, the entire Novartis collection of 1.5 million compounds was docked, optimized, and the 120 magnet features extracted for all 16 protein structures of the UEKR. The docking was performed on 190 Opteron CPU's. Table 1 shows the CPU times required by the various steps. By far the most time-consuming step was minimization in Flo+. As a comparison, minimizing in MOE took a similar time. Overall this exercise took all of the unused cluster cycles for 5 months, but having been done, it can be used for every kinase model in the future.

Table 1. Docking Timings

step	CPU s/cpd	CPU y/1.5 M cpds	processed-poses/ cpd
Dockit/Magnet filtering	272	13	2400
Flo+ minimization and scoring	832	40	177
extract Magnet features on nonredundant pose set	1	0.04	8.6

Design of the 10 000 Compound Screening Set. A 10 000 compound kinase-focused screening set was designed to emphasize known or predicted kinase activity as well as chemical-space coverage. First, all screening compounds with strong experimental kinase activity (100 nM IC_{50} against any kinase) were clustered using Scitegic's FCFP6 descriptors in PipelinePilot[13] to 3047 compounds. This selection was combined with medium-active (5 μM IC_{50}) compounds and reclustered at a larger radius. Clusters lacking a representative from the primary selection each contributed a compound, adding a total of 3341 compounds to the screening set. In selecting a representative compound from a cluster, priority was given to compounds active against the kinases least frequently hit by previously selected compounds, with ties broken by minimum IC_{50} against any kinase.

In addition to experimental actives, predicted kinase actives were also sampled. A Bayesian model was trained on 45 095 experimental "kinaphiles" (sub 5 μM IC_{50}, or sub 2 μM logit estimation from a single-concentration screen) and 65 737 "kinaphobes", defined as compounds tested against at least 18 kinases without once showing kinase activity. While a true experimental verification of inactivity would (impractically) require testing against all 500+ protein kinases, using these kinaphobes was preferred over "random" compounds. Guided by the model's separation of known kinaphiles from kinaphobes, the top 93 933 of 595 617 "drug-like" predicted kinaphiles were clustered with the selected experimental actives at an even larger radius, adding 3920 final compounds for a total of 10 000.

RESULTS

Simple Surrogates. Compounds with pIC_{50} data for any of eight kinases were combined and docked into all eight crystal structures. For each kinase, plain AutoShim models were built to predict the activity from its native crystal structure, and surrogate AutoShim models were built from each of the other seven crystal structures. The lower eight rows of Table 2 show the full matrix of these results. The upper number of each cell holds (surrogate) AutoShim predictive-R^2 for 25% withheld data; the lower number is for the "unshimmed" Flo+ scoring. The eight plain AutoShim models are on the diagonal, and the 56 surrogate AutoShim models are off-diagonal. Rebuilding AutoShim models from randomized starting poses indicates reproducibility is about 0.03.

Due to the lack of prescribed hydrogen bonds without steric clashes, some crystal structures successfully dock fewer compounds.

Table 2. Correlation (R^2) with pIC_{50} for the 25% Withheld Test Set Using Shimmed[a] and Unshimmed[b] Models, for Surrogate (Off-Diagonal) and Native (Italic) Docking Models[c]

	target IC_{50} data set[e]							
crystal structure[d]	CSF1R	CHK1	PDK1	AurA	GSK3b	PI3Ka	TIE2	PIM1
ensemble[f]	[0.60]	[0.68]	[0.54]	0.50	0.43	[0.51]	0.57	0.20
	[0.24]	[0.14]	[0.05]	[0.14]	0.03	0.08	0.05	0.03
CSF1R	***0.56***	0.54	0.44	0.37	0.37	0.40	**[0.64]**	0.19
	0.23	0.01	0.01	0.05	0.00	−0.00	−0.00	0.04
CHK1	**0.56**	***0.64***	0.47	0.44	0.38	**0.50**	0.54	0.19
	0.09	*0.10*	0.03	0.09	0.00	0.00	−0.04	**[0.06]**
PDK1	0.49	0.62	*0.49*	0.48	0.40	0.39	0.51	**[0.25]**
	0.09	**0.13**	***[0.05]***	0.08	0.01	0.00	−0.02	0.03
AurA	0.49	0.62	0.49	***0.56***	0.35	0.41	0.61	0.17
	0.12	0.03	0.02	***0.11***	0.00	0.05	0.16	0.02
GSK	0.50	0.57	**0.53**	0.50	*0.39*	0.43	0.59	0.17
	0.05	0.10	0.03	0.10	*0.01*	0.07	**[0.20]**	0.03
PI3K	0.47	0.57	0.40	0.40	0.31	*0.40*	0.49	0.13
	0.07	0.02	0.02	0.10	0.00	***[0.10]***	0.02	0.01
Tie2	0.35	0.47	0.39	0.41	0.34	0.31	*0.24*	0.20
	0.00	0.00	0.00	0.02	0.01	0.02	*0.00*	0.01
PIM1	0.42	0.57	0.45	0.39	**[0.45]**	0.38	0.27	*0.15*
	0.07	0.04	0.01	0.06	**[0.11]**	0.00	0.09	*0.02*

[a] The top value listed for each crystal structure is AutoShim R^2. [b] The bottom value for each structure is unshimmed Flo + R^2. The Flo+ score is defined for each compound as the highest score across all docked poses for a particular structure, or across all 16 structures for the ensemble case. [c] For each IC_{50} data set, the best shimmed and unshimmed result is bracketed. The best result exclusive of ensemble is in bold. [d] Rows name the crystal structure used for docking. The average number of compounds that successfully dock into that structure across data sets orders rows. [e] Columns name the IC_{50} data set for training and testing. [f] The 16-structure universal kinase ensemble receptor.

The top row of Table 2 lists results for ensemble surrogate AutoShim models trained on the UEKR model of 16 active sites. The ensemble Flo+ score for a particular compound is the highest score across all 16 UEKR docking models. The overall best kinase docking model for each target's data set is placed in square brackets, with the best excluding the ensemble model in bold. Counting square brackets in the top row, the ensemble model outperforms any single surrogate for four of the eight targets (one in nine would be expected by chance), and the correlation is within 0.07 of the best single-surrogate model for every kinase. The ensemble model also has the advantage that more compounds can be modeled. To escape prediction, all 16 docking models must eliminate a compound.

Excluding the ensemble AutoShim results from analysis, native AutoShim (on the diagonal) performs better than any single surrogate, or ties for best, in three of eight cases (vs one in eight expected). This native preference highlights AutoShim's sensitivity to protein features particular to each native kinase structure. In other words, AutoShim is performing something more than just 3D-QSAR.

Results for the all-purpose Flo+ scoring function, both for native, surrogate and UEKR docking, demonstrate the importance of nonsequence-related factors in docking accuracy. While the "native" structure outperforms all simple surrogates in four of eight cases, the best Flo+ score across the UEKR predicts activity as well or better than the native structure (on the diagonal) in all cases except PI3K, which is also the only lipid kinase in the ensemble. The other ensemble members are all protein kinases, with corresponding low active site homologies to PI3K (see Table 3).

Ensemble Surrogate AutoShim. The usage of multifeature shims derived from RP was investigated out of concern that rules composed of many terms might hit so few compounds as to promote overfit models. While trees are terminated after seven splits, resulting in a maximum of seven features in any terminal node, not all branches reach this depth. The actual distribution of the number of features in multifeature shims is plotted in Figure 2a. The distribution peaks at 4, excluding the single features extracted from the "seventh tree". Figure 2b shows that among the test poses selected by AutoShim (always the highest scoring), the more complex multifeature shims are about equally likely to match poses as the simpler shims. This could be explained by the fixed complexity parameter in rpart, which halts a branch's partitioning if further splits cannot improve the fit by a threshold value. This limits generating multifeature shims targeting an exceptionally small pose population. The total number of descriptors feeding PLS for the UEKR AutoShim models ranges from 18 to 397, averaging 196.

Table 3. Active Site Residue Percent Identity Matrix[a]

	CSF1R	CHK1	PDK1	AurA	GSK3	PI3K	TIE2	PIM1
CSF1R	100	62.5	55	52.5	51.3	15	52.5	40
CHK1	62.5	100	55	52.5	48.7	10	52.5	45
PDK1	55	55	100	50	48.7	17.5	50	37.5
AurA	52.5	52.5	50	100	53.8	17.5	57.5	47.5
GSK3	50	47.5	47.5	52.5	100	22.5	50	47.5
PI3K	15	10	17.5	17.5	23.1	100	20	22.5
TIE2	52.5	52.5	50	57.5	51.3	20	100	47.5
PIM1	40	45	37.5	47.5	48.7	22.5	47.5	100

[a] The number of identical residues divided by the total number of residues in the top-row's active site sequence defines percent identity. For each protein, active site residues are those within 4.5 A of the crystallized ligand. However, only the 40 sequence-aligned residues (39 for GSK3) defined as an active site residue in at least two structures were used for the identity calculation.

Profile Boost. Predicting the activity against any single kinase target with a linear combination of surrogate ensemble AutoShim models, parametrized by PLS, improves the median over the individual models by a modest 0.02 R^2-units. This is far less improvement than was observed on the same data sets for 2D-QSAR models.[15]

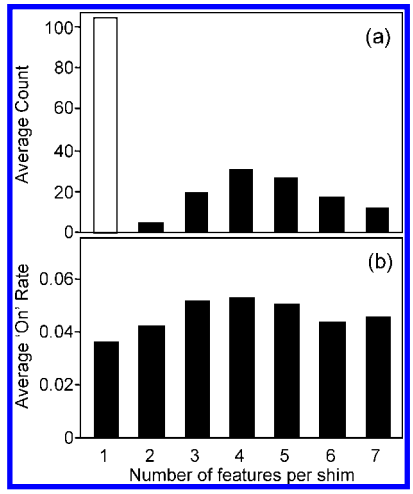

Figure 2. Shim complexity and on-rate. Statistics on the number of component features in a shim extracted from RP that contribute to final AutoShim models, limited to the 41 "successful" AutoShim models for which R^2 against the test set exceeds 0.32. (a) Average number of shims containing a given number of features across all 41 models. Terminal nodes contribute almost no single-term rules (0.2 per model, on average). The unfilled bar at 1 reports the average count of single-feature shims from gathering contributing features of multifeature shims in the seventh tree (see Methods). (b) Average frequency any rule turns on a bit in the best-scoring (per compound) test set poses, by depth. The single feature shims from the seventh-tree are excluded from this calculation.

Determinants of good models. Ensemble surrogate AutoShim models were generated for 68 kinases. Of these, 56 do not contribute a crystal structure to the UEKR and are therefore purely ensemble *surrogate* AutoShim models. Results for these 56 data sets evaluate the quantitative accuracy that can be achieved by a target specific scoring function *without* a crystal structure. Figure 3 shows the predictive R^2 for the 68 kinase targets, as assessed by correlation on 25% randomly withheld data. Kinases that contribute to the protein structure ensemble are identifiable in Figure 3 by having Flo+ performance plotted. Activity correlations vary widely, from 0.04 to 0.71. In principle, this variability might be due to features of the target, or features of the surrogate structures.

Figure 4 shows the performance of (nonensemble) surrogate AutoShim using 8 different protein structures as the surrogate for each of 68 kinase targets. One might expect that performance would depend on the homology of the surrogate. Strikingly, the quality of the results depends mainly on the target, not on the particular choice of surrogate protein structure. This indicates that the variability in Figure 3 noted above is not due to the surrogate structures, but still might be due to features of the target itself, features of the training data set, or the quality of the assay data.

Figure 5 plots profile-boosted ensemble AutoShim prediction performance against dynamic range as measured by the standard deviation of the IC_{50} data for the 68 kinase targets. Correlation between training set pIC_{50} standard deviation and AutoShim predictive R^2 is 0.53. This suggests that data sets must contain sufficient numbers of active compounds to train effective shims. The triangular point cloud indicates that wide dynamic range assures good correlations, but good models are sometimes obtained with limited dynamic range. Red color, which concentrates above the regression line, indicates the advantage of more training set compounds. Combined, these two criteria sensibly suggest that good models require a sufficient number of active compounds. The data sets with associated crystal structures are marked as triangles in Figure 5. While Figure 3 might suggest that having a crystal structure contributing to the ensemble improves performance, as most of these lie on the higher R^2 half of Figure 3, Figure 5 tells us that two of these lie well above the regression line, indicating an advantage from real structural data, but otherwise the superior performance coincides with large dynamic range. Based on these data, we determined that the surrogate AutoShim models work best if the training sets include at least 40 sub-micromolar hits. 51 of the 68 kinases meet this criterion.

Table 4 shows that across these 51 ensemble surrogate AutoShim models with at least 40 sub-micromolar actives, the median predictive R^2 on the withheld test set is 0.53 and the maximum is 0.71. For the 12 targets that contribute an X-ray structure to the ensemble, the median is slightly higher at 0.56, than for the 39 pure surrogate AutoShim models without a crystal structure at 0.52. Relative to docking with a general scoring function, for which the median R^2 is 0.03 and the maximum is 0.24, ensemble AutoShim offers a huge improvement. Overall, 80% (41) of the 51 ensemble surrogate AutoShim models exceed $R^2 = 0.32$, our criterion for success, which was the best performance for conventional docking with a native crystal structure, in the nearly 300 attempts from the GSK study cited in the introduction.[1]

Tests against 50 000 PDK1 High-Throughput IC_{50} Values. A 3D surrogate AutoShim model was trained against 7750 high-quality PDK1 IC_{50} values. The model was used to predict (retrospectively) previously measured high-throughput, 4-concentration, nonduplicated, PDK1 IC_{50} values of 49 145 additional compounds, to test the performance of the models on a large data set. Figure 6 plots the experimental vs. the surrogate AutoShim predicted activity. Lines are drawn at 1 μM, dividing the plot into quadrants: true and false positives and negatives. The correlation is $R^2 = 0.5$. An additional 2264 compounds from the lower-quality screening experiment that were among those in the high-quality training experiment serve as a control. The two experiments only correlated with $R^2 = 0.6$. Since the model was trained to reproduce the high-quality training data, this is the upper limit that an ideal model would achieve, i.e. the observed correlation of $R^2 = 0.5$ was out of a possible $R^2 = 0.6$.

Predictions were also made with a 2D profile-QSAR model trained on the same data (see below). The Venn diagram in Figure 7 details the agreement between the screening experiment and the 2D and 3D models. Table 5 summarizes the performance for the 2D profile-QSAR and 3D AutoShim models, as well as for the consensus and union of the two.

DISCUSSION

Fast 3D Modeling on Precomputed Dockings. One hurdle to practical application of docking for virtual screening is the time and computational expense of docking millions of compounds into each new protein target of interest. Because the UEKR can be shimmed to reproduce the activity data for any kinase target, large molecule collections can be predocked into just these 16 structures, and the stored magnet features from these dockings can be used for any of the

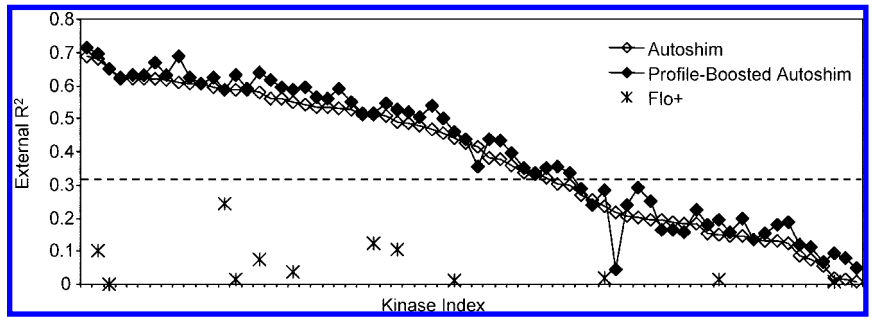

Figure 3. Predictive R^2: Flo+, AutoShim, profile-boosted AutoShim. For reference, the docking success criterion of $R^2 = 0.32$ is drawn.

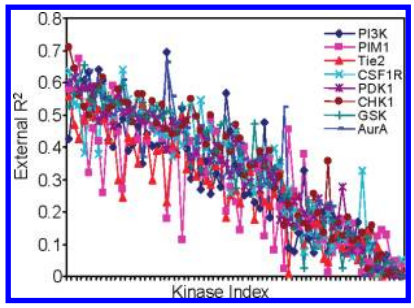

Figure 4. Simple (nonensemble) surrogate AutoShim. Crystal structures for eight kinases serve as surrogate docking models for AutoShim-training on 68 kinase activity data sets.

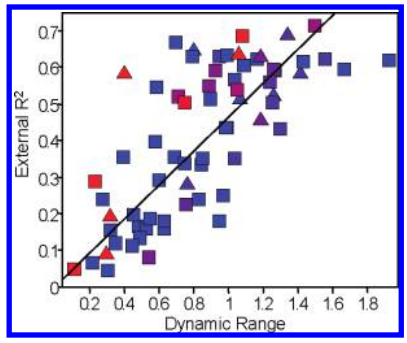

Figure 5. AutoShim R^2 dependence on training set dynamic range. Profile-boosted AutoShim performance on the external test data set is plotted against dynamic range of the training data, defined by the standard deviation in pIC_{50} values. Values for each of 68 targets are color coded by the number of compounds in the training set (few [blue] to many [red]). Targets that contribute a crystal structure have triangle symbols. Pure-surrogate models are squares. The best-fit line, forced through the origin, is also plotted.

remaining >500 kinases. Toward that end, the Novartis screening collection of 1.5 million compounds was docked into the 16 protein structures. The ligand geometries were minimized in Flo+, the poses filtered for redundancy, and the 120 atom count and H-bond distance features for each pose were extracted with Magnet and stored. This extensive calculation took all of the spare cycles on a cluster for several months, but can now be used for every kinase target.

Now, for each new kinase of interest, with IC_{50} training data, only a simple series of relatively quick steps remain to build and test a target-customized scoring function by iterative RP-on-PLS/pose selection: (1) run RP on the training data to define the optimal single- and multipoint shims, (2) compare them to the stored vector of 120 magnet features for each training compound pose to generate the shim-fingerprints, (3) combine these with the Flo+ scores for each pose to train a PLS model, iterating with pose selection, and finally, (4) test the model on a 25% held-out test set. This

Table 4. Flo+, Ensemble AutoShim, Profile-Boosted AutoShim, and 2D-QSAR[a]

	kinase set[b]	range	median	mean
Flo+	12 UEKR	0.00–0.24	0.03	0.06
unboosted AutoShim	all 51	0.01–0.69	0.51	0.45
	12 UEKR	0.02–0.68	0.53	0.46
	39 non-UEKR	0.01–0.69	0.49	0.45
profile-boosted AutoShim	all 51	0.08–0.71	0.53	0.48
	12 UEKR	0.09–0.69	0.56	0.49
	39 non-UEKR	0.08–0.71	0.52	0.47
2D-PLS	all 51	0.19–0.82	0.62	0.57
	12 UEKR	0.25–0.74	0.63	0.57
	39 non-UEKR	0.19–0.82	0.61	0.57

[a] 2D-QSAR performed with PipelinePilot[13] using PLS on FCFP6 descriptors. [b] Performance statistics limited to the 51 kinases with at least 40 actives at micromolar affinity. The 12 UEKR members that contribute an IC_{50} data set.

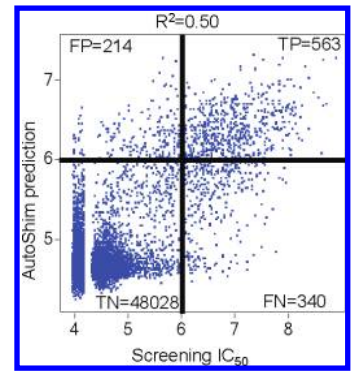

Figure 6. Surrogate AutoShim predicted IC_{50} vs screening IC_{50} for ~50 000 compounds, divided into quadrants at 1μM, indicating true and false positives and negatives.The correlation is $R^2 = 0.5$, but this probably underestimates the predictive power of the model, because the screening IC_{50} values do not correlate well with the high-quality 8-concentration IC_{50}s ($R^2 = 0.6$).

takes less than a day. If the model is predictive, two more very quick steps remain to perform virtual screening: (5) generate the shim-fingerprints from the RP model from the stored vector of 120 magnet features for each nonredundant pose of the compound archive and (6) run each of these through the PLS equation to predict the activity of each compound in the Novartis Archive. Even for the 1.5 million compound screening collection, spread over a 190 CPU cluster this only takes about $^1/_2$ h. The following flowchart summarizes the steps that are performed just once on the entire corporate archive and the steps that are performed for every new kinase target:

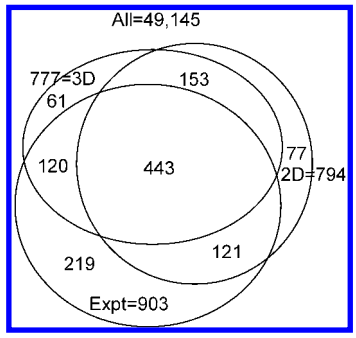

Figure 7. Venn diagram showing common and unique hits between 2D profile-QSAR predictions, 3D surrogate AutoShim predictions, and HTS experimental actives at 1 μM. Results are summarized in Table 5.

Table 5. Performance against PDK1 Screening Data, at 1 μM, of a 3D Surrogate AutoShim Model, a 2D Profile-QSAR, and the Consensus and Union of the Two

prediction method	true neg	true pos	false neg	false pos	hit rate[a]	recovery rate[b]	enrichment
3D AutoShim	48028	563	340	214	72%	62%	39X
2D QSAR	48012	564	339	230	71%	62%	39X
consensus[c]	48089	443	460	153	74%	49%	40X
3D/2D union[d]	47951	684	219	291	70%	76%	38X

[a] Hit rate = 100 × TP/(TP + FP). [b] Recovery rate = 100 × TP/(TP + FN). [c] Consensus = actives predicted by both the 2D and 3D method. [d] Union = actives predicted by either the 2D or 3D method.

Perform once...

(1) Select and superimpose 16 X-ray structures.

(2) Cluster ligand atoms to define 20 atom count and 30 H-bond shim centers.

(3) Dock 1.5 million cpds into 16 X-ray structures (months on cluster).

(4) Filter poses to <10 representatives/compound.

(5) Extract 120 atom count and H-bond distance Magnet "features" for each pose.

For each new kinase target...

(1) MTS assay at least 400 compounds with at least 40 sub-micromolar actives.

(2) Split into 75% training and 25% test set.

(3) Run 5 iterations of PLS-on-RP on training set ($^1/_2$ day).

(a) RP on Magnet features of current best poses to build multifeature shims with count/distance thresholds and extract corresponding fingerprints for all poses.

(b) Run PLS on shim fingerprints of current best poses to get shim-weights.

(c) Predict new best poses.

(4) Test model against 25% test set.

(5) Compare final shims with the predocked magnet-feature database to generate target-specific fingerprints for each pose of the 1.5 million cpds.

(6) PLS predictions on 1.5 million predocked compounds (1 h).

Proposed Application of Iterative Screening. Surrogate ensemble docking has so far been used for scaffold hopping in mature projects, where sufficient IC_{50} data were available to train the shimmed models from lead optimization campaigns, and for following up high-throughput screening (HTS) hits. Given that HTS can take over half of a year and cost a million dollars, a bigger potential benefit would come from virtual screening tools that could find important hits in advance of HTS. Newly available cherry picking sample-management robots allow for new virtual screening-based hit-finding paradigms. The empirically trained ensemble surrogate AutoShim scoring functions, however, require initial training data, and cannot be used de novo. Toward that end, an iterative screening process has been developed that starts with an initial screen of a carefully designed set of 10 000 kinaphilic compounds (see Methods). This training set emphasizes diverse known kinase inhibitors, diverse predicted kinase inhibitors, diverse selectivity patterns, and representation of the structural diversity of the Novartis screening collection.

These 10 000 compounds are screened in duplicate at 25 μM, and 8 concentration IC_{50}s are determined for the hits. Because the compounds have already been predocked into the 16 kinases of the UEKR, and their features extracted with magnet, building the optimal shims and extracting shim-fingerprints with RP takes under an hour. Iterative PLS/pose selection on 7500 compounds takes about $^1/_2$ day, and testing on the remaining 2500 withheld test compounds takes minutes. If the model is predictive, virtual screening of the 1.5 million compounds on the precomputed magnet features takes less than an hour. The structures, properties, and predicted poses of the active compounds are then examined (aided by interactive use of Magnet), and several thousand of the most attractive compounds will be retrieved by cherry-picking robots and tested. Besides producing hits, adding any false positives to the training data will improve the model for additional iterative rounds of virtual screening.

Not As Quantitative As 2D-QSAR, but Potentially Better for Scaffold Hopping. While ensemble surrogate AutoShim is nearly as fast as 2D-QSAR on the predocked archive and is highly predictive compared to docking with a general purpose scoring function, it is still slightly less predictive than our best kinase-specific 2D-QSAR method. We have developed kinase specific in-house 2D profile-QSAR models that combine IC_{50} data from over 100 000 compounds tested across nearly 100 kinases in the activity prediction of each new kinase target.[15] Like surrogate AutoShim, this very highly predictive method required a large initial investment (in experiments), but is nearly as fast to apply as conventional 2D-QSAR. Table 4 shows that across the 51 kinases in this study, the median R^2 on a 25% held out test set using our 2D profile-QSAR models was 0.62 and the maximum was 0.82 compared to a median of 0.53 and a maximum of 0.71 using ensemble surrogate AutoShim. Why, then, would one bother with the less predictive 3D method? 2D-QSAR models rely on topological similarity. Conventional wisdom says that docking models and 3D-QSAR, which do not "know about" 2D topology, work equally well on structurally dissimilar actives topologically unrelated to the training set, i.e., docking can make larger "scaffold hops", albeit at the cost of lower accuracy and correspondingly more false positives. There is little direct literature support for this generally held belief, and it is difficult even to imagine a systematic study that would confirm it. However, a recent paper that compiled "examples of successful scaffold hops" indirectly supports it.[14] Only 2 of the 21 examples in their survey of successful scaffold hops used strictly 2D methods. This is the rationale behind conventional docking, which barely correlates with activity.

AutoShim finally provides a simple and general 3D method that approaches the best 2D-QSAR and should be that much more useful for scaffold hopping.

While we have not yet proved experimentally that surrogate AutoShim is more efficient at scaffold-hopping than 2D profile-QSAR, we can at least show that the relatively orthogonal 3D method can find different actives than the 2D profile-QSAR method. The test against PDK1 screening data at 1 μM in Table 5 shows that the performance of the 3D and 2D methods was virtually identical in this case, with hit rates ~72%, recovery rates ~62%, and enrichments ~40-fold. (Since the screening experiment correlated only poorly with the high-quality IC_{50}s, $R^2 = 0.6$, see above, much of the misclassification must be due to the screening experiment itself, rather than the predictions.) Consensus scoring had a marginally higher hit rate, but a substantially lower recovery of actives, so it was not beneficial in this case. However, testing the union of predictions from both models finds 120 additional actives over either model alone, substantially raising the recovery rate to 76%, but still with a 70% hit rate. Also, the 153 false positives from the consensus model are particularly likely candidates for experimental false negatives that might be recovered. Unfortunately, too few actives at 1 μM in this set would qualify as completely "new scaffolds", so conclusions about scaffold hopping could not be drawn.

CONCLUSION

For the kinase family, surrogate ensemble AutoShim has addressed the three chief limitations of conventional docking: (1) it does not require a target protein structure, (2) its predictions correlate with measured affinity, and (3) it is as fast as 2D-QSAR (at least for the predocked 1.5 million compound Novartis screening collection). It obviates the need for a structure of the target protein by using a UEKR made up of 16 very diverse protein structures. It achieves its predictive power by using training IC_{50} data to shim the active site with pharmacophore-like interaction features that compensate for errors in the general purpose scoring function and for the use of surrogate protein structures. It achieves its speed by predocking our entire compound archive into the universal receptor, reducing a docking and scoring problem to simple algebra.

While surrogate ensemble docking has so far only been used for scaffold hopping in mature projects, a much bigger potential benefit will come from an iterative screening process in advance of HTS. This approach will combine medium-throughput screening of an initial designed set of 10 000 kinaphiles, followed by model building and virtual screening on the predocked 1.5 million compound screening archive, followed by compound retrieval and testing.

Surrogate AutoShim has so far been applied only to the kinase family of proteins. This family was an attractive starting point with a highly conserved and well defined ATP binding site and a large number of potential therapeutic kinase targets. Kinases are also challenging, due to the large active site conformational flexibility. We are optimistic that the method will work equally well with other therapeutically important protein families such as proteases, phosphodiesterases, and nuclear hormone receptors.

ABBREVIATIONS:

UEKR, universal ensemble kinase receptor; RP, recursive partitioning; PLS, partial least-squares; HBA, hydrogen bond acceptor; HBD, hydrogen bond donor; HTS, high-throughput screening; MTS, medium-throughput screening.

REFERENCES AND NOTES

(1) Warren, G. L.; Andrews, C. W.; Capelli, A. M.; Clarke, B.; Lalonde, J.; et al. A Critical Assessment of Docking Programs and Scoring Functions. *J. Med. Chem.* **2006**, *49*, 5912–5931.
(2) 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.
(3) *Magnet*, version 1; Metaphorics LLC: Santa Fe, NM, 2001.
(4) Oshiro, C.; Bradley, E. K.; Eksterowicz, J.; Evensen, E.; Lamb, M. L. Performance of 3D-Database Molecular Docking Studies into Homology Models. *J. Med. Chem.* **2004**, *47*, 764–767.
(5) Kairys, V.; Fernandes, M. X.; Gilson, M. K. Screening Drug-Like Compounds by Docking to Homology Models: A Systematic Study. *J. Chem. Inf. Model.* **2006**, *46*, 365–379.
(6) Lesk, A. M.; Chothia, C. Mechanisms of domain closure in proteins. *J. Mol. Biol.* **1984**, *174*, 175–191.
(7) Janin, J.; Rodier, F. Protein-protein interaction at crystal contacts. *Proteins: Struct., Funct., Genet.* **1995**, *23*, 580–587.
(8) Diller, D. J.; Li, R. Kinases, Homology Models, and High Throughput Docking. *J. Med. Chem.* **2003**, *46*, 4638–4647.
(9) Ferrara, P.; Jacoby, E. Evaluation of the utility of homology models in high throughput docking. *J. Mol. Model.* **2007**, *13*, 897–905.
(10) *DockIt*, version 1.5; Metaphorics LLC: Santa Fe, NM, 2001.
(11) McMartin, C.; Bohacek, R. S. QXP: Powerful, rapid computer algorithms for structure-based drug design. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 333–344.
(12) Ajay, A. On better generalization by combining two or more models: a quantitative structure-activity relationship example using neural networks. *Chemom. Intell. Lab. Syst.* **1994**, *24*, 19–30.
(13) *Pipeline Pilot*, version 6.0; Scitegic: San Diego, CA, 2006.
(14) Schneider, G.; Schneider, P.; Renner, S. Scaffold-Hopping: How Far Can You Jump. *QSAR Comb. Sci.* **2006**, *25*, 1162–1171.
(15) Martin, E. J.; Sullivan, D. C. Kinase-family profile-QSAR modeling. *J. Chem. Inf. Model.*, in preparation.

CI700455U