JOURNAL OF
CHEMICAL INFORMATION
AND MODELING

ARTICLE

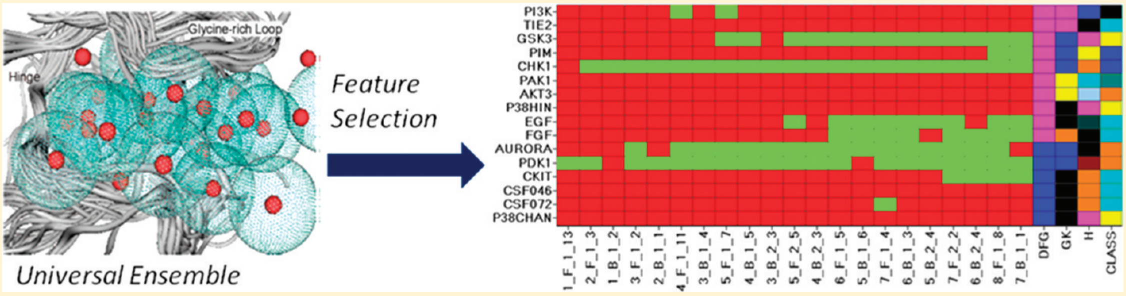
pubs.acs.org/jcim

Development of a Minimal Kinase Ensemble Receptor (MKER) for Surrogate AutoShim

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

ABSTRACT:



The target-tailored 3-D virtual screening (VS) method "Surrogate AutoShim" adds pharmacophoric shims to a 16-kinase crystal structure "Universal Kinase Ensemble Receptor" (UKER) to generate highly predictive, target-customized docking models. Predocking a corporate archive of millions of compounds into the 16-structure ensemble takes months. However, since the 16 UKER structures are always the same, docking need only be done once. The predocked results are then "shimmed" to reproduce experimental training data for any number of additional kinases far more accurately than conventional docking. Training new kinase models and predicting activity for millions of predocked compounds against dozens of kinases takes only hours. However reducing the predocking time would make the method even more advantageous. Sequential Floating Forward Search (SFFS) was employed to rationally identify a reduced subset using only 8 of the 16 structures, a "Minimal Kinase Ensemble Receptor" (MKER) that preserved the predictive accuracy for 20 kinase models. Furthermore, a performance evaluation of this subset on an extended set of 52 kinase targets and 100,000 compounds showed statistical model performance comparable to the original UKER. The MKER has halved the time for predocking large databases of internal and commercial compounds. For *ad hoc* virtual libraries, where predocking is not possible, 2- or 3-kinases "Approximate Kinase Ensemble Receptors" (AKER) were also identified with only a modest loss of prediction accuracy.

INTRODUCTION

Surrogate AutoShim[1,2] is a kinase-specific, target-tailored, virtual screening (VS) technique (Figure 1) that lies between conventional docking and 3D QSAR. It models inhibition of any given kinase target by "shimming" a fixed surrogate "Universal Kinase Ensemble Receptor" (UKER) of diverse kinase X-ray structures to reproduce experimental assay training data. The "shims" are pharmacophore interactions which are combined with the docking energy and weighted via PLS regression to give a target-customized scoring function used to predict the activity of additional compounds. The original UKER, into which each compound is docked, consisted of 16 diverse X-ray structures from 14 kinases. They were selected to approximate the full range of shapes and pharmacophore arrangements available to any given new kinase. The structures and some of their features are detailed in Table 1. Up to 100 poses per compound are obtained for each of the 16 kinases, and 9 diverse poses are selected to sample this larger set. The specific protein in each ensemble docking is treated simply as part of the "pose", and the multiple pose problem[3] is solved through an iterative cycle of pose selection and scoring-function optimization, which converges in just 2 to 4 iterations. The shims fill in the additional target-specific details required for accurate activity prediction.

Since the same 16-kinase UKER is used to model any of the remaining 500 kinases, each compound need only be docked into the 16 UKER structures once. The "raw" docking score and pharmacophoric interactions are saved in a database and reused to build each subsequent Surrogate AutoShim model. Although model building and scoring are very fast, the initial predocking of a large compound collection is slow. Docking a corporate screening archive of ~1.8 million compounds into the 16 UKER structures took 8 months using all the spare cycles on a cluster of about 200 processers. By contrast, building quantitative models for 52 kinases based on experimental training data, followed by predicting activity for the ~1.8 million compound archive on all 52 kinase targets, took only 2 days.

Received: May 25, 2011
Published: August 17, 2011

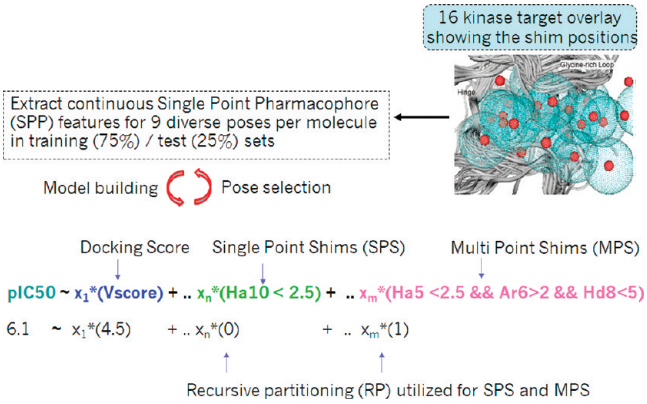

16 kinase target overlay showing the shim positions
Glycine-rich Loop
Hinge
Extract continuous Single Point Pharmacophore (SPP) features for 9 diverse poses per molecule in training (75%) / test (25%) sets
Model building
Pose selection
Docking Score
Single Point Shims (SPS)
Multi Point Shims (MPS)
pIC50 ~ x1*(Vscore) + .. xn*(Ha10 < 2.5) + .. xm*(Ha5 <2.5 && Ar6>2 && Hd8<5)
6.1 ~ x1*(4.5) + .. xn*(0) + .. xm*(1)
Recursive partitioning (RP) utilized for SPS and MPS

Figure 1. Workflow for Surrogate AutoShim model generation. Training compounds are docked into the UKER, 9 diverse poses are selected, and their pharmacophoric interactions are stored. These interactions are added as "shims" to the scoring function and weighted by PLS regression, to reproduce experimental IC_{50} training data. New best poses are selected by the revised scoring function, and training is iteratively repeated, converging in 2 to 4 cycles.

Table 1. Classification of the Structures Used in the UKER

enzyme[g]	DFG_conf[a]	GK[b]	SB_αC[c]	SB_DFG[d]	HINGE[e]	CLASS[f]
akt3	IN	M	N	N	EYV	AGC
aurora	OUT	L	N	N	EYA	AGC
chk1	IN	L	Y	Y	EYC	CAMK
ckit	OUT	T	Y	N	EYC	TK
csf046	OUT	T	N	N	EYC	TK
csf072	OUT	T	Y	N	EYC	TK
egf	IN	T	N	Y	QLM	TK
fgf	IN	V	Y	N	EYA	TK
gsk3	IN	L	Y	Y	DYV	CMGC
p38chan	OUT	T	Y	N	HLM	CMGC
p38hin	IN	T	Y	N	HLM	CMGC
pak1	IN	M	N	N	EYL	STE
pdk1	OUT	L	Y	N	SYA	AGC
pi3k	IN	I	N	N	EIV	LIPID
pim	IN	L	Y	Y	ERP	CAMK
tie2	OUT	I	N	N	EYA	TK

[a] DFG_conf — IN or DFG-in and OUT or DFG-out. [b] GK — gatekeeper. [c] SB_αC — salt bridge between catalytic Lysine and αC acidic residue. [d] SB_DFG — salt bridge between catalytic Lysine and Asp from DFG. [e] HINGE — sequence of GK+1 to GK+3 residues. [f] CLASS — subfamily to which the kinase belongs. [g] Note that csf046/csf072 and p38chan/p38hin are different conformations of the same kinase.

Surrogate AutoShim has 3 big advantages over conventional docking: it is far more accurate, does not require a new protein structure, and for predocked compounds it is extremely fast. However, it has 2 disadvantages: it requires training data, and initially predocking large compound databases into the 16 structure UKER is slow. The 16 structures were chosen by "art" to span the widest possible range of pocket geometries and pharmacophore arrangements for the UKER. The selection was intentionally generous; no attempt was made to minimize the number of structures in the ensemble. Kinase computational and structural chemistry experts within the Novartis community were canvassed for structures presenting unusual conformations, residue types, or ligand poses. These were visually compared to the 8 structures used in the original (nonsurrogate) AutoShim validation.[4] 8 additional structures were selected to augment the target family and structural diversity.

A validation study reported in the Surrogate AutoShim paper[4] using the 8 kinases with both structures in the UKER and IC_{50} data sets for modeling showed that, in 7 out of 8 cases, the models built from the UKER outperformed models using the native kinase structures confirming the quality of the selected UKER.

While there were no obvious redundancies, it was anticipated that the ensemble was overengineered, and that due to multicolinearities, a smaller subset of the 16 kinase structures might do just as well as the full set. The time required for predocking would be reduced proportionally. This becomes important as the method is extended beyond the corporate archive to even larger collections of commercially available compounds and to virtual libraries. An effort was therefore made to find a minimal spanning subset of the kinases that would give models with the predictive power of the full UKER but with the fewest kinase structures. Identifying a Minimal Kinase Ensemble Recepter (MKER) is a feature selection problem. Recognized in the fields of artificial intelligence[5] and machine learning,[6] feature selection algorithms select a minimal subset of features (i.e., a subset of the 16 UKER kinase X-ray structures) that can still reproduce an objective function. A number of algorithms are available, depending upon the size and complexity of the feature set, candidate set, and objective function. The methodologies may be primarily classified into two groups: "filter" and "wrapper" methods.

Filter methods try to find the most relevant feature subset *a priori*, without considering the effects on a target function. One approach evaluates the correlation of features with the dependent variable and eliminates variables with low correlation. Another approach identifies and eliminates highly correlated variables from a descriptor matrix prior to QSAR model generation. However, these methods only deal with linear dependencies and may eliminate critical information for nonlinear methods such as Recursive Partitioning (RP), Random Forests (RF), and Support Vector Machines (SVM). Furthermore, very high correlation or anticorrelation between variables does not always lead to a lack of complementarity between variables, and variables which by themselves hold little correlation with the dependent variable, when combined with others, might still be useful in improving the overall prediction quality.[6]

In the wrapper approach,[5] the feature selection algorithm exists as a wrapper around an "induction algorithm", treating the later as a black box. The wrapper generates a trial subset of features. The induction algorithm then uses this feature set to train a model and make predictions on a test set, to evaluate the quality based on the objective function. The results are fed back to the wrapper, which executes a step of decision making, and iteratively selects the next feature subset for evaluation. The process continues in an iterative fashion. The termination criteria may be a certain number of iterative cycles, a desired subset size, or the achievement of a desired end point based on the objective function. While one can perform a brute force exhaustive search using all possible combinations, the problem is NP-hard and quickly becomes prohibitively expensive with increasing numbers of features or for complex objective functions. Several methods have therefore been developed to conduct this search in a thorough yet time efficient manner.

Wrappers may be divided into two categories: deterministic and stochastic. The oldest deterministic wrapper is the Sequential Forward Search (SFS),[7] which simply progresses through the addition of the new feature to the existing feature set that results in

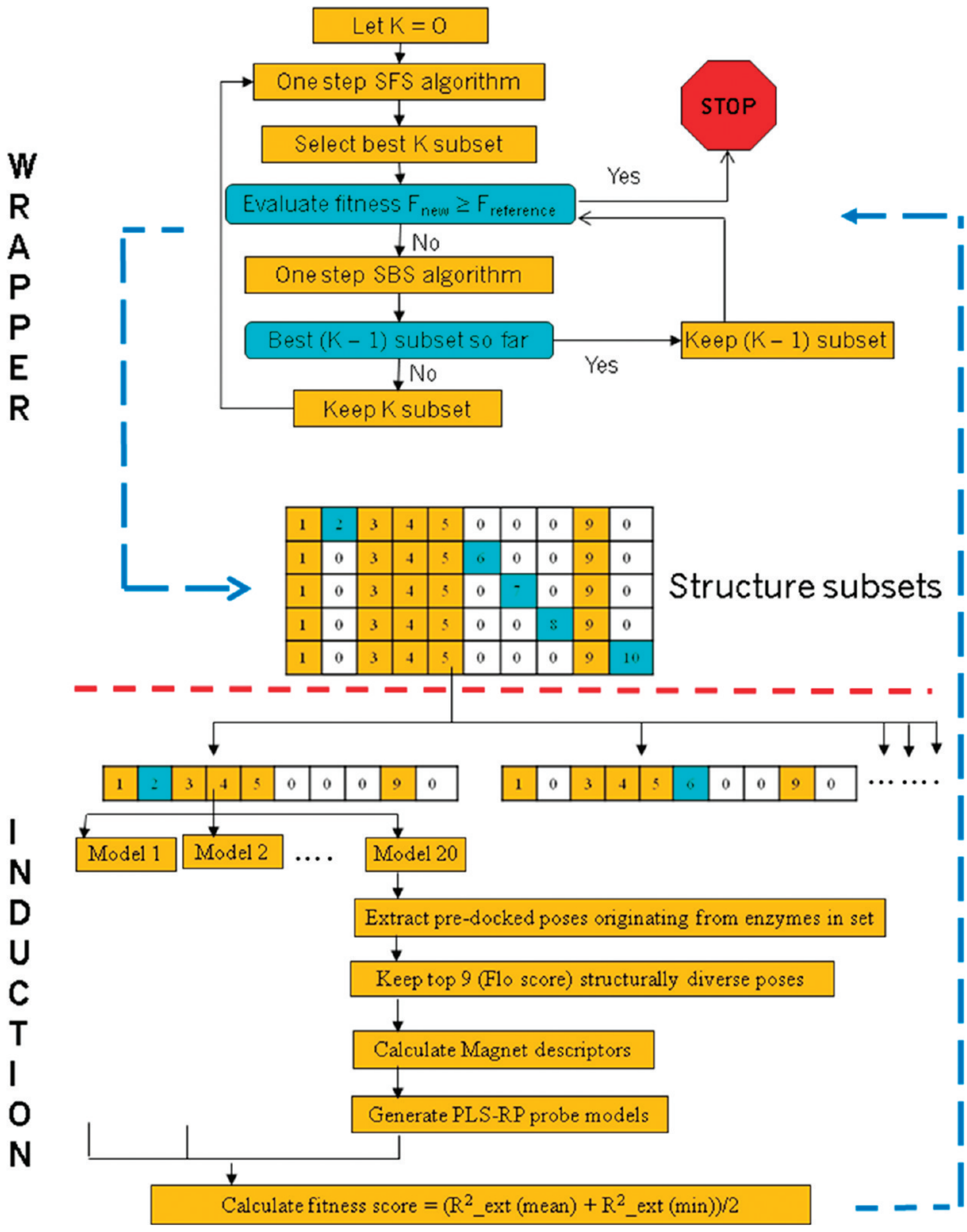


Figure 2. A workflow showing the implementation of the SFFS search for identifying putative MKERs.

the largest improvement of the objective function. Sequential Backward Search (SBS)[8] is the opposite of SFS. It starts with a full feature set and carries out sequential elimination of features which preserves or improves the objective function. Both of these methods suffer from "nesting", which for SFS would refer to the inability of the algorithm to remove a previously added feature, even if the elimination would lead to the improvement of the objective function. Sequential Floating Forward/Backward Search (SFFS/SFBS)[9] is a more complex variant of the SFS/SBS which involves additional backtracking steps after each addition/deletion to evaluate eliminations/additions that might aid in an overall improvement of the objective function. In this way, the algorithm corrects for premature inclusion/deletion decisions during its progression and circumvents the nesting problem. Adaptive Sequential Floating Forward/Backward Search (ASFFS/ASSFBS) is a more complex variant of the SFFS/SBBS which allows for a degree of generalization as the current subset size approaches the desired subset size. However,[10] in later studies the authors have noted that compared to SFFS/SBBS, the adaptive version brings a 1−5% improvement only in some cases and at a large cost of computational overhead. Oscillating search (OS)[11] may be considered as a cross between a deterministic and a stochastic wrapper, wherein feature subsets of varying size (depending upon the algorithm criteria) are added to or removed from the existing subset during each step of the search progression. Among the wrapper methods described so far, this may be the only one that is applicable to large sets of 10000+ features.

Stochastic wrappers are typically designed for large search spaces. They try to overcome the NP-hard problem by incorporating a degree of randomness into the search. Genetic Algorithms (GA)[12] emulate the biological process of evolution. A "population" of feature subsets (chromosomes) are generated in each iteration and evaluated based on a "fitness function". The best feature sets are "selected" and then have part of their features swapped (crossover) or replaced (mutation) to generate the next

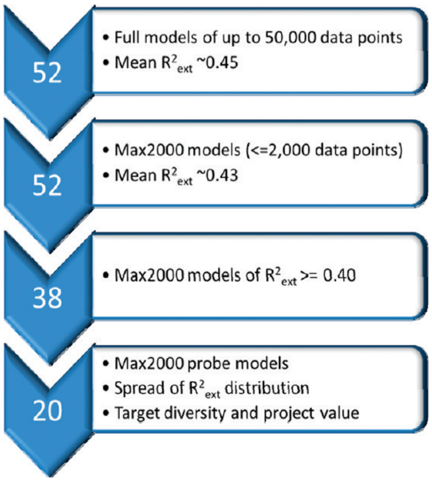

52
• Full models of up to 50,000 data points
• Mean R^2_{ext} ~0.45
52
• Max2000 models (<=2,000 data points)
• Mean R^2_{ext} ~0.43
38
• Max2000 models of R^2_{ext} >= 0.40
20
• Max2000 probe models
• Spread of R^2_{ext} distribution
• Target diversity and project value

Figure 3. Selection of the reduced "probe model" set of kinase targets and compounds for the SFFS search.

population of feature subsets. The algorithm executes over a large number of cycles and converges toward the desired fitness criterion. GA has been used extensively,[13] including QSAR[14] and docking.[15,16] In Simulated Annealing (SA)[17] the algorithm probabilistically determines whether to move the system in a state S^1 to a new randomly perturbed state S^2 according to probabilities chosen, such that over a set of iterations the system converges toward the optimal fitness criterion. This method has been used extensively in molecular mechanics and QSAR-based roles such as ADAPT.[18] Other algorithms include swarm optimization[19,20] and ant-colony optimization.[21]

The feature set in this case was the 16 crystal structures that form the Universal Kinase Ensemble. The induction algorithm used to create the objective function summarizes the statistical quality of Surrogate AutoShim models generated from the docked poses originating from a particular subset of the 16 crystal structures. The induction algorithm is a very slow step in each iterative cycle of algorithmic feature selection. Stochastic wrapper methods are designed for larger feature sets and take more iterations to converge. Therefore, a deterministic wrapper method was selected. The SFFS/SFBS method shows a significant improvement over the more simplistic SFS/SBS methods which are prone to "nesting" and also performs very near or equally well to the more complex and slower ASFFS/ASFBS methods. Several comparative studies have shown that SFFS produces desirable results, at a fraction of computational costs, compared to more computationally expensive methods. Ferri et al.[22] compared SFS, SFFS, and GA on data sets of up to 360 features. They showed that the performance of GA and SFFS were similar on smaller data sets, but the GA performance degraded with increasing data set size. Kudo et al. compared SFFS with GA[23] and found that the former is better suited to small and medium scale problems, while the latter is more applicable to large scale problems. Jain et al.[24] likewise compared SFFS with several other methods and found it to be an excellent performer in terms of accuracy and speed. Therefore, the SFFS feature selection algorithm was chosen to determine the MKER.

■ METHODS

Resources. A 160 core Linux blade center (each blade having a 2.8 GHz quad core CPU and 16 GB of memory) was used for all calculations. All scripting was carried out using Python 2.4 and shell scripting was utilized for parallelization.

Workflow. The detailed implementation of the SFFS search is flowcharted in Figure 2. The section above the red dotted line shows the wrapper component of the protocol. For example, say the current best MKER candidate consists of $K = 6$ structures. The method adds each of the remaining 10 structures to those 6 generating 10 $K = 7$ structure MKER candidates, which are then passed on to the induction phase (below the red dotted line). In the standard Surrogate AutoShim workflow for the full UKER, a set of up to 100 poses per protein structure have been previously generated for all compounds by docking with Dockit,[25] followed by molecular mechanics minimization in Flo+.[26] In each SFFS cycle, the standard Surrogate AutoShim workflow is followed, except that for each molecule in each of 20 probe kinase assay data sets, only the subset of the 1600 predocked poses are retrieved that originated from the structures in the current candidate MKER. The usual Surrogate AutoShim procedure is then followed: 1) Up to nine representative poses are selected to optimize the Flo+ score and conformational diversity. 2) Magnet[27] extracts the standard set of 122 pharmacophore interactions descriptors for each pose. 3) The shim set and training IC_{50} data are input to the model building R script, which implements Surrogate AutoShim's "Partial Least Squares on Recursive Partitioning" (PLS-RP) procedure to generate single- and multipoint shims based on RP of the pharmacophore descriptors. 4) These, in addition to the Flo+ docking score, are used as independent variables to generate the PLS model. 5) IC_{50}s are predicted for the 25% held-out test set, and the correlation R^2_{ext} is computed as the figure of merit.

Upon completion of the induction phase, the final fitness score for each candidate MKER was calculated from the average and minimum R^2_{ext} for the 20 kinases and passed back to the wrapper. Fitness scores are calculated for all the structure subsets, and the current best fitness score (F_{new}) is compared against the reference fitness score ($F_{reference}$), i.e. the score obtained for the UKER. If $F_{new} > F_{reference}$, then the search is stopped. If not, the search is continued with a backtracking step of SBS on the winning MKER candidate of $K = 7$ structures, and the new best $K = 6$ structure subset thus selected is compared to the best previous subset of $K = 6$ structures to check for the "nesting" problem. The iterations continue until $F_{new} > F_{reference}$, or the full set of all 16 structures is reached.

Implementation of SFFS Search. The SFFS algorithm described in the original paper[9] was implemented in Python 2.4. The script takes a list of all previously run subsets of kinase structures and the best subset from the current run (based on the fitness score) as inputs. It outputs the next set of subsets (i.e., candidate MKERs), for evaluation in the induction phase. The induction phase was coded to take advantage of parallel processing across a cluster of about 200 nodes, such that each core would execute the steps starting from pose retrieval to model training and testing for a given kinase model with a given structure subset. For example, a SFFS iteration involving a population of 10 candidate MKERs, each applied to 20 kinase models (each model taking 1 h) would correspond to 200 model generations and would be completed in 1 h on the 200 node cluster. The pbs scripts submit the Surrogate AutoShim pose extraction, descriptor calculation, and model building job creation as well as IC_{50} predictions on the held-out test sets. Previously published Magnet .sea scripts and R scripts[28] were utilized for shim calculation and model building. Since the MKER must be a subset of the UKER, the search used predocked

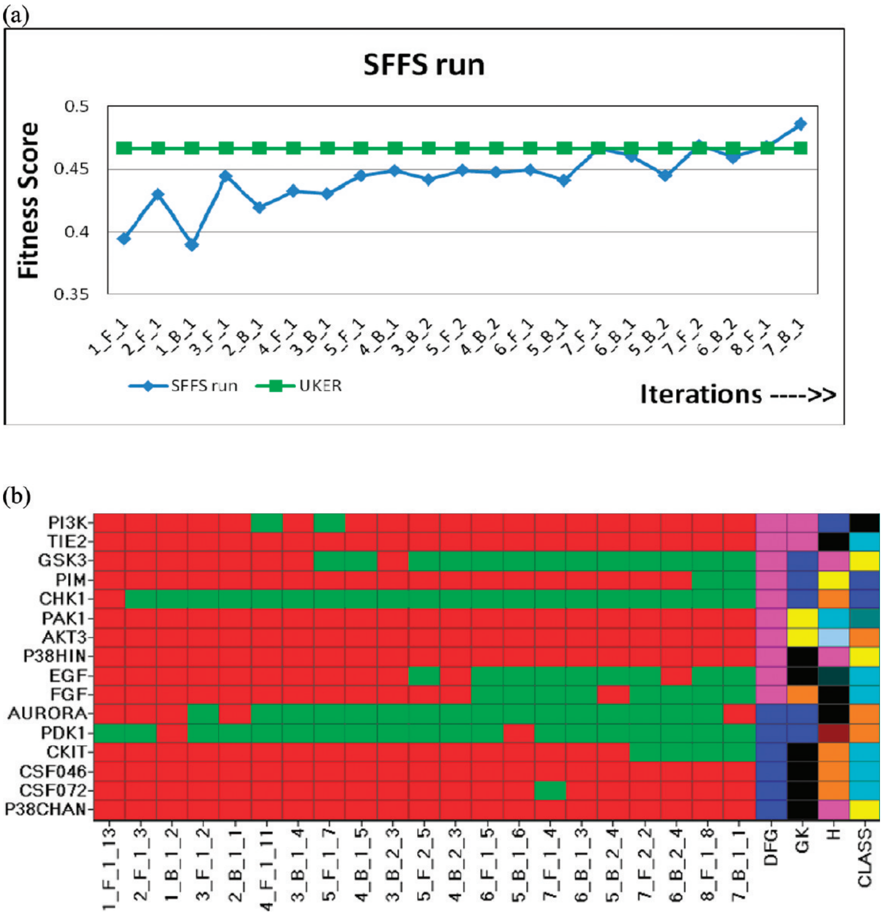


Figure 4. (**a**) A plot showing the progression of the feature selection process. The Fitness score is plotted on the *Y*-axis, while the sequential steps of the selection are plotted on the *X*-axis. The blue line shows the fitness scores obtained from the search, while the green line shows the corresponding fitness score obtained by training the 20 "Max2000" probe models on the full UKER. The step names, e.g. 1_F_1_13, fit a descriptive convention: 1 = a 1 feature subset, F = using an SFS Forward selection step, 1 = first attempt at finding a 1 feature subset using SFS, 13 = the 13th subset from the generated subsets (out of a possible 16) had the highest fitness score. (**b**) The specific kinase structures employed from the full set of 16 structures at each step of the selection process are in green. On the far right the structures are categorized by 4 criteria as detailed in Table 1: 1) "DFG" — DFG loop conformation, 2) "GK" — gatekeeper residue, 3) "H" — hinge sequence, and 4) "CLASS" — subfamily of protein kinase or lipid kinase.

Table 2. Four Candidate MKERs Selected Based on SFFS and the Frequency of Each Kinase among the Four Candidates

7_B_1_1	7_F_2_2	7_F_1_4	8_F_1_8	frequency
	aurora	aurora	aurora	3
chk1	chk1	chk1	chk1	4
ckit	ckit		ckit	3
		csf720		1
egf	egf	egf	egf	4
fgf	fgf	fgf	fgf	4
gsk3	gsk3	gsk3	gsk3	4
pdk1	pdk1	pdk1	pdk1	4
pim			pim	2

Table 3. Comparing the Mean R^2_{ext}, Minimum R^2_{ext}, Fitness Score, and Mean R^2_{train} from the Full UKER to the 4 Candidate MKERs and 2 AKERs Modeling All 52 Kinases with the Full Data Sets

ensemble	R^2_{ext}(mean)	R^2_{ext}(min)	fitness score	R^2_{train}(mean)
UKER	0.455	0.016	0.236	0.618
7_B_1_1	0.448	0.04	0.244	0.629
7_F_2_2	0.442	0.031	0.237	0.631
7_F_1_4	0.453	0.047	0.25	0.633
8_F_1_8	0.439	0.028	0.234	0.632
3_F_1_2	0.422	0.059	0.24	0.624
2_F_1_3	0.415	0.024	0.219	0.621

poses of compounds originating from the corporate archive, with proprietary internal biological data for model training and testing, so no additional docking needed to be done. Theoretically, these phases could be run seamlessly in an iterative fashion. However, periodic problems with cluster submissions, such as jobs refused by the slave nodes or glitches in communication with the head node, cause a small number of jobs to fail, so error checking and resubmission of failed jobs is needed at the end of each induction step. To avoid extensive defensive programming, the trivial step of data transfer between the wrapper and the induction phase at the end of each cycle was done by hand.

Study Design. The "gold standard" benchmark, against which all subsequent analyses would be compared, was the cumulative performance of 52 kinase models (Figure 3), spanning the kinome, built from the docked poses in the full 16-structure UKER. Each assay data set was divided into 75% training and 25% held-out test sets. The squared correlation coefficient between the predicted and experimental values for the 25% held-out test set (R^2_{ext}) was the measure of quality for each Surrogate AutoShim model. The 52 kinases included 2 nonhuman kinases and 50 human kinases (including mutants), of which 6 were lipid kinases and the remainder were protein kinases. The size of the kinase assay data sets ranged from 600 IC_{50}s for early stage projects to upward of 50,000 IC_{50}s for kinases used in a large profiling study. A typical application of Surrogate AutoShim would be in early hit finding, where the iterative model building for smaller training data sets (up to a few thousand compounds) could be done in a few hours. However, model building takes 1−2 days for the largest data sets and would be prohibitively slow for each search iteration. Therefore, working "probe" kinase assay data sets were curtailed to a maximum of 2000 IC_{50}s, so all model building runs for a given iteration would finish relatively quickly. Assays with fewer than 2000 IC_{50}s were carried forward unchanged. For larger assay data sets, 2000 compounds were selected at random. The 52 models rebuilt on these "Max2000" data sets had an average R^2_{ext} of ~0.43 which was comparable to an average R^2_{ext} of ~0.45 for the full data sets. Furthermore, a set of just 20 "probe" kinases was selected from the original 52 data sets to reduce the computational cycle time for each SFFS iteration. For this purpose, all Max2000 data sets with an R^2_{ext} < 0.40 were eliminated, and from the remainder, the set of 20 "probe models" was chosen based on a spread of R^2_{ext} values, target diversity and project interest. The objective function for the SFFS, shown in eq 1, was designed to reward both high average performance and high worst-case performance. The SFFS search termination criterion for a candidate MKER was set at a fitness score = 0.46 for the 20 "Max2000" probe models, the same score achieved using all 16 structures in the UKER

$$Fitness\_Score = \frac{\overline{R^2_{ext}} + \min(R^2{}_{ext})}{2} \qquad (1)$$

where

$$\overline{R^2_{ext}} = Mean\ value\ from\ 20\ models$$

$$\min(R^2_{ext}) = minimum\ value\ from\ 20\ models$$

RESULTS AND DISCUSSION

Results from SFFS Search. Figure 4 summarizes the progress of the SFFS search in finding a MKER. The PDK1 DFG-out structure with an SYA hinge from the AGC family is the overall best single surrogate (Figure 4b) for all 20 kinases, with a fitness score of 0.39 (Figure 4a). Large fluctuations were observed in the first few search steps, where each new structure can add a lot of new binding-site features. The method picks up CHK1, a DFG-in structure from the CAMK family with a EYC hinge sequence and AURORA a DFG-"out-like" structure from the AGC family with a EYA hinge sequence, as the next most informative kinase structures, raising the fitness score to 0.45. The addition of the fourth, fifth, and the sixth structures do not show much increase in the fitness score. However, the addition of the seventh and the eighth structures do increase the fitness score to 0.46 which is comparable to that obtained from the UKER. More back tracking steps are observed beyond the addition of the third structure. The search was allowed to proceed for a few more iterations to address nesting issues and to identify other candidate MKERs of similar predictive power. In the first SFS step, when all 16 structures are in the unused pool, evaluating 20 probe data sets requires building 20*16 = 320 Surrogate AutoShim models. Since the first SBS step has only 2 potential candidate structures to remove, it requires only 40 models. As the candidate MKER set grows, the number of models needed for each SFS step decreases, and the number for each SBS step increases. A total of 3600 Surrogate AutoShim models were required in order to identify the final candidate MKERs of 7 or 8 structures. This suggests that the structural diversity of the kinase ATP binding site is in some sense 7- or 8-dimensional.

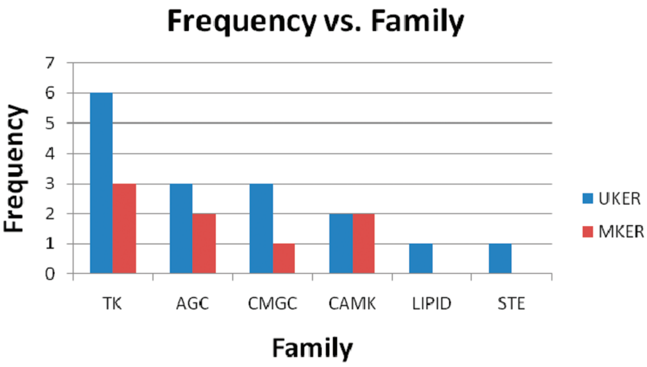


Figure 5. A distribution of the kinase structures from the UKER and MKER based on their family associations.

Analysis of Optimized Subsets. Four candidate MKERs (Table 2) from the SFFS search were selected for further consideration. Since they had been optimized only on the probe subset of 20 models and reduced compound sets, performance still needed to be confirmed on the full set of 52 kinases, using the full data sets not clipped to a maximum of 2000 points. Table 3 compares the mean R^2_{ext}, minimum R^2_{ext}, fitness score, and mean R^2_{train} (on the 75% training set) from the full UKER to the 4 candidate MKERs and 2 AKERs modeling all 52 kinases with the full data sets. The fitness scores from all 4 candidate MKERs were comparable to the original UKER. The structures from the STE and LIPID families did not occur in any of four candidate MKERs. Five structures appeared in all 4 candidate MKERs: 2 AGC kinase structures (chk1, pdk1), 2 TK kinase structures (egf, fgf), and 1 CMGC kinase structure (gsk3). The "8_F_1_8" ensemble of 8 structures was selected based on the diversity of structural features. This ensemble had 3 structures (out of 7 from UKER) with DFG-out and 5 structures (out of 9 from UKER) with DFG-in. It also included Pim, which has a less common proline in the hinge, thereby eliminating one hydrogen bonding site. Figure 5 shows the distribution of the UKER and MKER structures based on their family associations. The two frequency distributions are well correlated (R^2 = 0.72). Although LIPID kinase structures did not appear in the MKER, the median R^2_{ext} for 7 LIPID kinase models was 0.39 in the MKER vs 0.40 in the UKER when tested with full compound data sets, suggesting that the LIPID kinases share enough similarity in their binding sites to protein kinases to be coded for within the protein kinase ensemble. There were only 3 STE kinases among the 52 models, with median R^2_{ext} = 0.24 in the MKER vs 0.26 in the UKER.

While the goal was to reproduce the full accuracy of the 16 structure UKER, 2-and 3-kinase "Approximate Kinase Ensemble

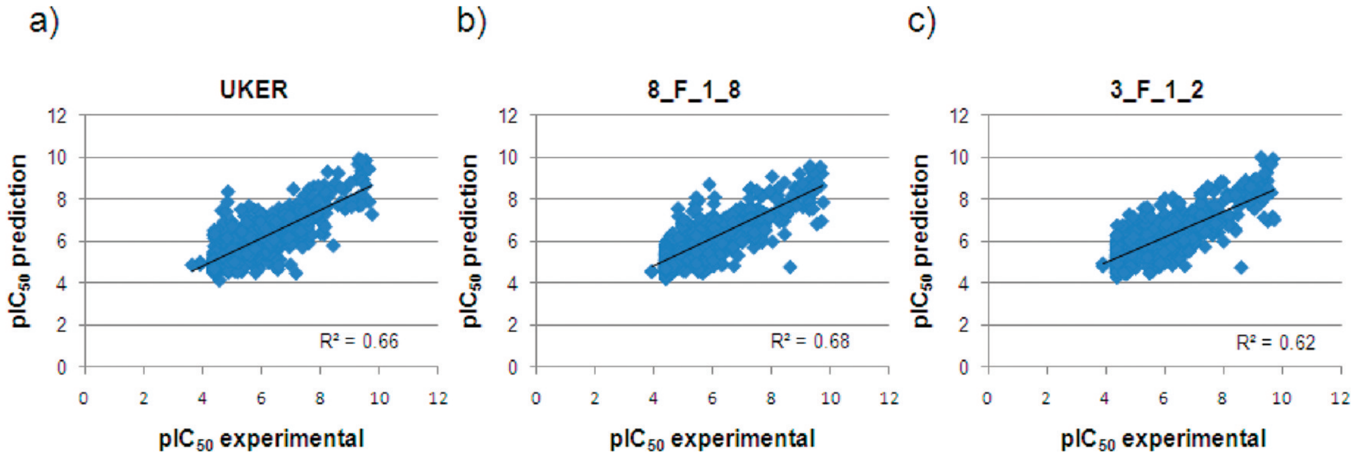


Figure 6. Experimental vs predicted pIC$_{50}$ plots obtained for the 25% held-out test set of an individual kinase using the (a) UKER, (b) MKER(8_F_1_8), and (c) AKER (3_F_1_2).

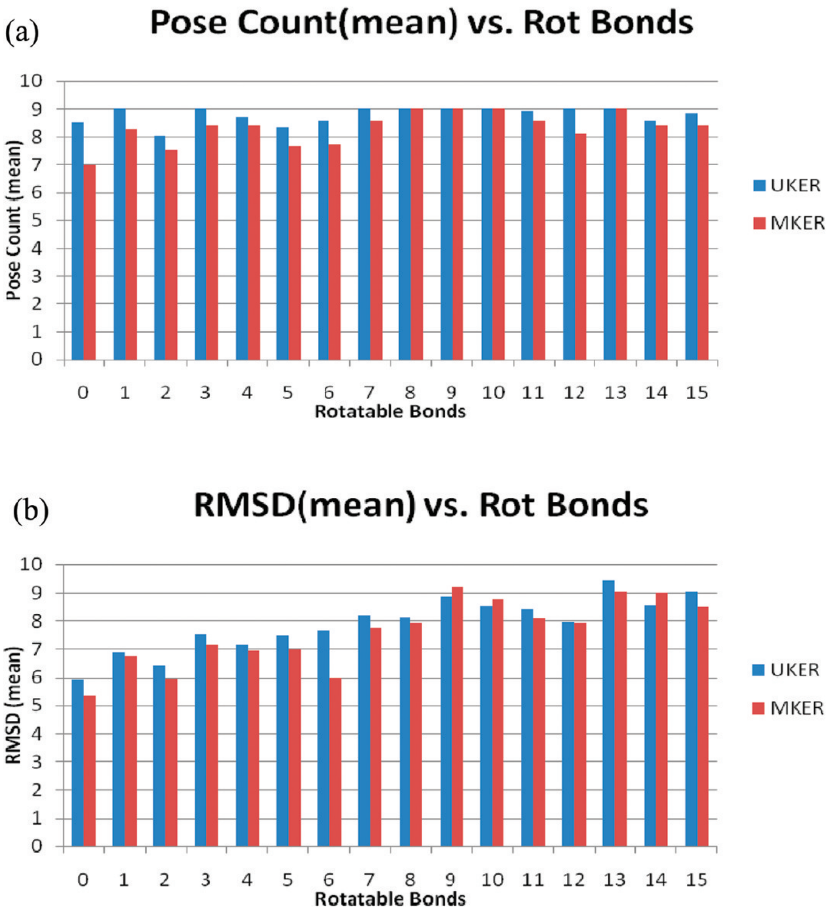


Figure 7. (a) The mean number of poses per ligand, binned by the number of rotatable bonds. (b) Pose diversity by number of rotatable bonds, calculated as the mean of the RMSDs between each ligand's top scored pose based on Flo+ docking score, and the up to 8 additional diverse poses.

Receptors" (AKER) (Table 3) achieved correlations of 0.42, compared to 0.46 for the UKER. For many applications where predocking is not available, such as docking *ad hoc* virtual libraries, the time saved docking into this even smaller subset would more than compensate for the modest loss of accuracy.

Figure 6 shows the experimental vs predicted pIC$_{50}$ plots for the 25% held out test set of an individual kinase obtained using the UKER, MKER (8_F_1_8), and AKER (3_F_1_2). The R^2_{ext} values for UKER and MKER are very comparable while that of the AKER is slightly less.

Evaluation of Pose Diversity. Dealing with the multiple-pose problem with an iterative training procedure is an important feature of Surrogate AutoShim model building. The target-customized scoring function is optimized, while simultaneously selecting the low-energy pose according to that function, from up to 9 diverse pose representatives algorithmically selected to sample the hundreds produced by docking.[1,2] The robust model training/pose selection procedure converges in just 2−4 cycles without overtraining. The diversity of available poses in the MKER was compared to the original UKER by evaluating the number of diverse representative poses found (up to 9), and the rmsd among the representative poses, grouped by rotatable bond (RB) count. A sampling of 240 ligands was evaluated: 16 sets of 15 ligands each, comprised of 0 to 15 rotatable bonds. Figure 7 (a) shows the mean number of diverse poses selected by the Surrogate AutoShim procedure.[4] The mean number of poses selected at each RB bin for the smaller ensemble of structures in the MKER is slightly lower than the larger UKER ensemble, but the difference is less than 1 pose except for RB = 0. The corresponding plot in Figure 7 (b) shows the mean rmsd between the top scored pose (based on Flo+ docking score) and the (up to 8) subsequent poses, averaged over each RB group. Mean rmsd and RB count were correlated, with R^2: 0.80 for MKER and 0.71 for UKER, indicating that more flexible ligands explore more conformational space. The mean rmsd for the UKER and MKER across the RB bins are similar, with the MKER trailing slightly in most of the cases.

As expected, the overall docking time using the 8-structure MKER ensemble was about 1/2 that of the full UKER.

CONCLUSIONS

The SFFS feature selection algorithm efficiently identified a suitable 8-structure subset of the original 16-structure UKER used by Surrogate AutoShim for 3D predictive modeling of kinase affinity. Surrogate AutoShim is a 3D protein-family based target-tailored scoring function generation and virtual screening method that predocks large databases of molecules into a diverse ensemble of crystal structures chosen to sample the structural diversity of a protein family. It builds predictive models by weighting pharmacophoric shims in the UKER along with molecular mechanics-based docking scores, to reproduce experimental IC_{50} training data. After predocking, models for 52 kinases were trained and accurate activity profiles for 1.8 million compounds computed in just 2 days. The predocking need only be done once, but still it took 8 months to predock the 1.8 million compounds. Decreasing the number of structures in the ensemble decreases the predocking time proportionally. The original UKER used for developing Surrogate AutoShim was an art-based selection designed to represent the widest array of conformational and feature diversity. While it predicted affinity very well, it had structural redundancies and/or multicolinearities that could be eliminated, decreasing the predocking burden without sacrificing accuracy.

SFFS is a deterministic, wrapper-based, feature-selection algorithm that converges rapidly for small to medium data sets. The fitness function optimized the average and minimum predictive accuracy of a set of 20 kinase probe models. Four candidate MKERs of only 7−8 structures were identified. They were further evaluated on an extended set of 52 models built on full data sets of up to 50,000 data points. All 4 candidate MKERs performed comparably to the original UKER. The final 8-structure MKER was selected based on the structural and feature diversity of the members. Besides comparable predictive power, the MKER sampled nearly the full binding pose diversity of the UKER. Two- or 3-structure Approximate Kinase Ensemble Receptors (AKER) were also identified for *ad hoc* predictions on compound sets that have not been predocked.

As expected, predocking into the 8-structure MKER took 1/2 the time of the full 16-member UKER. Since Surrogate AutoShim prediction is very fast compared to predocking, the time to get full 52-kinase profiles for a collection of compounds is likewise halved. Predocking a database of ∼1.3 million commercially available compounds with the MKER took about 3 months. For *ad hoc* virtual libraries, which do not benefit from predocking, using 2- or 3-kinase AKERs would cut the time to make predictions of 50,000 compounds from about a week with the UKER to about a day, with only a modest loss in predictive accuracy.

As the Surrogate AutoShim methodology is extended toward other protein families, this search methodology would be prospectively utilized in conjunction with art-based selection to enhance the orthogonality of the ensemble members and improve accuracy and efficiency. Unlike the kinase application, where specific hinge hydrogen bond constraints filter out many poses prior to energy minimization, a larger number of poses may have to be minimized. Hence, a rational reduction in the ensemble size will be even more valuable.

Structural biologists and computational chemists involved in pharmaceutical drug discovery are often faced with decisions on the relevance of protein structures showing unique side-chain/backbone movements and their scope in terms of explaining SAR of specific series of compounds. While this is often an art-based judgment done at a detailed level with value for a given project, it might also be interesting to gauge their relevance on a more global and/or protein family level. The current implementation provides a tool for evaluation and selection of protein structures based on their ability to explain the variance in experimental data across multiple chemotypes and multiple targets.

AUTHOR INFORMATION

Corresponding Author

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

ACKNOWLEDGMENT

P.M. would like to thank the NIBR Education Office for Post Doctoral funding.

REFERENCES

(1) 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.

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

(3) Dong, X.; Ebalunode, J. O.; Cho, S. J.; Zheng, W. A Novel Structure-Based Multimode QSAR Method Affords Predictive Models for Phosphodiesterase Inhibitors. *J. Chem. Inf. Model.* **2010**, *50*, 240–250.

(4) 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.

(5) Kohavi, R.; John, G. H. Wrappers for feature subset selection. *Artificial Intelligence* **1997**, *97*, 273–324.

(6) Guyon, I.; Elisseeff, A. An introduction to variable and feature selection. *J. Machine Learning Intell.* **2003**, *3*, 1157–1182.

(7) Whitney, A. W. A direct method of nonparametric measurement selection. *IEEE Trans. Comput.* **1971**, *20*, 1100–1103.

(8) Marill, T.; Green, D. M. On the effectiveness of receptors in recognition systems. *IEEE Trans. Inf. Theory* **2010**, *9*, 11–17.

(9) Pudil, P.; Novovicova, J.; Kittler, J. Floating search methods in feature selection. *Pattern Recognit. Lett.* **1994**, *15*, 1119–1125.

(10) http://ro.utia.cz/fs/fs_guideline.html (accessed 4/1/2011).

(11) Somol, P.; Pudil, P. Oscillating search algorithms for feature selection. *15th IAPR International Conference on Pattern Recognition, Barcelona* 2000, pp 406-409.

(12) Goldberg, D. E. *Genetic Algorithms in Search, Optimization and Machine Learning*; Kluwer Academic Publishers: 1989.

(13) Von, , H. A. Evolutionary Algorithms and their Applications in Chemistry. .*Handbook of Chemoinformatics*; 2010; pp 1239−1280.

(14) Rogers, D.; Hopfinger, A. J. Application of Genetic Function Approximation to Quantitative Structure-Activity Relationships and Quantitative Structure-Property Relationships. *J. Chem. Inf. Model.* **1994**, *34*, 854–866.

(15) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and validation of a genetic algorithm for flexible docking *J. Mol. Biol.* **1997**, *267*, 727–748.

(16) Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.* **1998**, *19*, 1639–1662.

(17) Kirkpatrick, R. G. C. D.; Vecchi, M. P. Optimization by Simulated Annealing. *Science. New Series* **2010**, *220* (4598), 671–680.

(18) Sutter, J. M.; Dixon, S. L.; Jurs, P. C. Automated Descriptor Selection for Quantitative Structure-Activity-Relationships Using Generalized Simulated Annealing. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 77–84.

(19) Agrafiotis, D. K.; Cedeno, W. Feature selection for structure-activity correlation using binary particle swarms. *J. Med. Chem.* **2002**, *45*, 1098–1107.

(20) Lin, W. Q.; Jiang, J. H.; Shen, Q.; Shen, G. L.; Yu, R. Q. Optimized block-wise variable combination by particle swarm optimization for partial least squares modeling in quantitative structure-activity relationship studies. *J. Chem. Inf. Model.* **2005**, *45*, 486–493.

(21) Shen, Q.; Jiang, J. H.; Tao, J. c.; Shen, G. L.; Yu, R. Q. Modified Ant Colony Optimization Algorithm for Variable Selection in QSAR Modeling: QSAR Studies of Cyclooxygenase Inhibitors. *J. Chem. Inf. Model.* **2005**, *45*, 1024–1029.

(22) Ferri, F.; Pudil, P.; Hatef, M.; Kittler, J. *Comparative study of techniques for large-scale feature selection*. In *Pattern Recognition in Practice IV*; Gelsema, E., Kanal, L., Eds.; 1994; pp 403-413.

(23) Kudo, M.; Sklansky, J. Comparison of algorithms that select features for pattern classifiers. *Pattern Recognit.* **2000**, *33*, 25–41.

(24) Anil, J. Feature Selection: Evaluation, Application, and Small Sample Performance. *IEEE Trans. Pattern Analysis Machine Intell.* **1997**, *19*, 153–158.

(25) DockIt. http://www.metaphorics.com/products/dockit.html (accessed 4/1/2011).

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

(27) Magnet. http://www.metaphorics.com/products/magnet/index.html (accessed 4/1/2011).

(28) Wehrens, R.; Mevik, B. H. pls: Partial Least Squares Regression (PLSR) and Principal Component Regression (PCR); R package version 1.2−0, 2006. http://www.cran.r-project.org/ (accessed 4/1/2010).