Technical Report

Department of Computer Science
and Engineering
University of Minnesota
4-192 EECS Building
200 Union Street SE
Minneapolis, MN 55455-0159 USA

TR 09-017

Affinity-based Structure-Activity-Relationship Models: Improving Structure-Activity-Relationship Models by Incorporating Activity Information from Related Targets

Xia Ning, Huzefa Rangwala, and George Karypis

May 29, 2009

Affinity-based Structure-Activity-Relationship Models: Improving Structure-Activity-Relationship Models by Incorporating Activity Information from Related Targets

Xia Ning[1], Huzefa Rangwala[2] and George Karypis[1]
[1]Department of Computer Science and Computer Engineering
University of Minnesota, Twin Cities, Minneapolis, MN 55455
[2]Department of Computer Science
George Mason University, Fairfax, VA 22030

May 18, 2009

Abstract

Structure-activity-relationship (SAR) models are used to inform and guide the iterative optimization of chemical leads, and play a fundamental role in modern drug discovery. In this paper we present a new class of methods for building SAR models, referred to as *affinity-based*, that utilize activity information from different targets. These methods first identify a set of targets that are *related* to the target under consideration and then they employ various machine-learning techniques that utilize activity information from these targets in order to build the desired SAR model. We developed different methods for identifying the set of related targets, which take into account the primary sequence of the targets or the structure of their ligands, and we also developed different machine learning techniques that were derived by using principles of semi-supervised learning, multi-task learning, and classifier ensembles. The comprehensive evaluation of these methods shows that they lead to considerable improvements over the standard SAR models that are based only on the ligands of the target under consideration. On a set of 117 protein targets obtained from PubChem, these affinity-based methods achieve an ROC score that is on the average 7.0%–7.2% higher than that achieved by the standard SAR models. Moreover, on a set of targets belonging to six protein families, the affinity-based methods outperform chemogenomics-based approaches by 4.33%.

1 Introduction

The pioneering work of Hansch *et al.*,[1,2] which demonstrated that the biological activity of a chemical compound can be mathematically expressed as a function of its physicochemical properties, led to the development of quantitative methods for modeling structure-activity relationships (SAR). Since that work, many different approaches have been developed for building such structure-activity-relationship (SAR) models.[3,4] These *in silico* models have become an essential tool for predicting the biological activity of a compound from its molecular structure and played a critical role in drug and chemical probe discovery by informing the initial screens, design, and optimization of chemical compounds with the desired biological properties.

Over the years, a number of methods have been developed for improving the accuracy of the SAR models that utilize additional information beyond the known ligands of the targets under consideration. One of the early methods utilizes approaches based on active learning and iteratively expands the set of training compounds used for learning the SAR models.[5] In this approach, the target's experimentally determined ligands are used to build an initial SVM-based SAR model. Compounds that are close to the decision boundary of the SVM model are then selected and treated as additional positive training examples for learning a new SVM model. This process is repeated multiple times until the performance of the learned model cannot be further improved. Probably the most widely used approaches for improving the quality of the SAR models are those based on chemogenomics.[6–8] The key idea behind these approaches is to

learn a SAR model for a family of proteins (e.g., GPCRs, Kinases, etc.) instead of an individual protein. The rationale of these approaches is that proteins belonging to the same protein family tend to bind to compounds that share certain common characteristics. Thus, by taking into account the known ligands from all the members of the family, better models can be learned. In these approaches, the single family-based model is trained using instances consisting of target-ligand pairs for all targets in the family and their ligands. This model can then determine the SAR score for a specific target and a specific compound by using it to predict that particular target-compound pair. The different chemogenomics-based approaches that have been developed differ on the features of the targets, compounds, and their complexes that they utilize (e.g., physicochemical properties,[9,10] protein structure,[11] amino acid sequence,[12] binding site descriptors,[13,14] topological descriptors,[9] protein-ligand fingerprints,[15] etc.), how they represent target-compound pairs (e.g., concatenation of descriptor vectors,[9] tensor products,[12] kernel fusion,[12] etc.), and the machine-learning methods that they use for learning the family-based models (e.g., support vector machines,[12,16] neural networks,[17] partial least-squares,[10,11,14] random forests,[15] multi-task learning,[16] etc.).

In this paper we present a different approach for improving the quality of SAR models that also utilizes activity information from other protein targets. This approach, referred to as *affinity based*, identifies a set of targets that are *related* to the target under consideration and then utilizes only activity information from these targets while learning the desired SAR model. Even though this approach shares some characteristics with those based on chemogenomics, its advantage is that, by using appropriate target-to-target similarity functions to identify the related targets, it can adapt to the characteristics of the protein target under consideration and lead to higher-quality SAR models. In addition, its adaptive nature allows it to select a smaller number of targets than those present in the entire family, or to select targets from different families if their use will lead to better quality models.

We developed and investigated different methods to identify the set of related targets and to incorporate their activity information into the affinity-based SAR model. Specifically, we developed different target-to-target similarity measures for identifying the set of related targets that take into account the primary structure of the targets themselves or the structure of their ligands. In addition, we developed three different machine-learning approaches for building the SAR models that were derived from the principles of semi-supervised learning,[18] multi-task learning,[19–22] and classifier ensembles.[23–25] The experimental evaluation of these methods on a set of 117 targets extracted from PubChem show that for nearly all of them, the incorporation of activity information from other targets leads to quality improvements in the resulting affinity-based SAR models. The best results are obtained for the ligand-based target-to-target similarity methods and the multi-task learning and classifier ensembles schemes, which achieve an average of 7.0%–7.2% ROC improvement. In addition, on a set of six protein families, the affinity-based methods achieve a 4.3% improvement over chemogenomics-based approaches.

2 Methods

2.1 Definitions and Notations

In this paper the protein targets and the compounds will be denoted by lower-case t and c characters, respectively, and subscripts will be used to denote specific targets and compounds. For each target t_i, its set of experimentally determined active compounds will be denoted by C_i^+, whereas its set of inactive compounds will be denoted by C_i^-. For reasons discussed later in Section 2.2, the set of compounds in C_i^- will be obtained by randomly sampling the compounds that do not belong in C_i^+. The entire set of targets under consideration will be denoted by $\mathbb{T}$ and the union of active compounds over all targets by $\mathbb{C}$ (i.e., $\mathbb{C} = \bigcup C_i^+$).

Each compound will be represented by a topological descriptor-based representation in which each compound is modeled as a frequency vector of certain subgraphs (descriptors) present in its molecular graph.[3] The similarity between two compounds c_x and c_y will be denoted by $\text{sim}_c(c_x, c_y)$ and will be computed as the Tanimoto coefficient of their descriptor-based representation.[26] The Tanimoto coefficient is given by:

$$\text{sim}_c(c_x, c_y) = \frac{\sum_k c_{x,k} c_{y,k}}{\sum_k c_{x,k}^2 + \sum_k c_{y,k}^2 - \sum_k c_{x,k} c_{y,k}}, \quad (1)$$

where k goes over all the dimensions of the descriptor space and $c_{x,k}$ is the number of times descriptor k occurs in compound c_x.

Given a compound c and a set of compounds C, the k most similar compounds (based on the Tanimoto coefficient) of c in C will be denoted by $\text{nbrs}_k(c, C)$ and will be referred to as c's k nearest-neighbor in C. For two sets of compounds C_x and C_y, $\text{Nbrs}_k(C_x, C_y)$ will denote the union of the k nearest-neighbors of each compound $c \in C_x$

in C_y, that is:

$$\text{Nbrs}_k(C_x, C_y) = \bigcup_{c \in C_x} \text{nbrs}_k(c, C_y). \quad (2)$$

Finally, to aid in the clear description of the different methods, we will use the term *specific target* to refer to the protein target for which we want to build a SAR model. Depending on the method, this SAR model will be built using either its own activity information (baseline SAR model) or additional information obtained from its set of related targets (affinity-based SAR model).

2.2 Baseline SAR Models

For each target t_i, we used support vector machines (SVM)[27] to build the baseline SAR model that relies only on t_i's own set of active compounds. Given a set of positive training instances $\mathcal{I}^+$ and a set of negative training instances $\mathcal{I}^-$, SVM learns a classification function $f(x)$ of the form

$$f(x) = \sum_{x_i \in \mathcal{I}^+} \lambda_i^+ \mathcal{K}(x, x_i) - \sum_{x_i \in \mathcal{I}^-} \lambda_i^- \mathcal{K}(x, x_i) \quad (3)$$

where λ_i^+ and λ_i^- are non-negative weights that are computed during training by maximizing a quadratic objective function, and $\mathcal{K}(.,.)$ is a *kernel* function that measures the similarity between the compounds. Given a new instance x, the sign of the prediction score $f(x)$ is used to determine the class of x. In addition, a set of compounds can be ranked based on their likelihood of being positive by sorting their prediction scores in non-increasing order.

In the context of our problems, the set of positive instances for t_i corresponds to its own set of experimentally determined ligands C_i^+. However, determining the set of compounds that will form the negative class is problematic for two reasons. First, in many target-ligand activity databases, only information about actual ligands of a particular target is available and information about non-binding compounds is not provided. Second, even when the activity information is obtained from screening assays, the negative information may not be very reliable as compounds can fail to bind to a particular target due to assay-related artifacts. Thus, the actual learning problem associated with building a SAR model is that of learning a classifier from only positive and unlabeled instances[28–31] (an instance is considered to be unlabeled if it is not positively labeled). An approach that has been successfully used in the past to address this problem is to select as negative instances a random subset of the unlabeled compounds.[4] Recent work has shown that under the assumption that the labeled instances are *selected completely at random*, the model learned from such randomly selected negative instances produces rankings that are equivalent to the real model.[31]

In this work, motivated by these empirical and theoretical results, the set of negative instances (C_i^-) for the baseline SAR model is obtained by selecting $|C_i^+|$ random compounds from $\mathbb{C} \setminus C_i^+$. This allows for the creation of equal-size positive and negative training sets. Moreover, by using $\mathbb{C} \setminus C_i^+$ as the pool of compounds from which to select the negative instances, it allows for the definition of a more realistic (and harder) negative class as it contains compounds that are known to bind to other protein targets. Note that the same C_i^- set is also used for defining the negative instances for all the affinity-based methods that are described in the subsequent sections.

2.3 Affinity-based SAR Models

In recent years, chemogenomics-based approaches have illustrated that the quality of the SAR models can be improved by taking into account the activity information of the proteins in the same family. However, the fundamental step in these approaches, which is building a SAR model based on all the proteins in the family, has a number of shortcomings. First, it can only be applied to protein families for which activity information is available for multiple members. Second, for a specific target t_i, the chemogenomics-based model may contain activity information from protein targets that may not be helpful for it (e.g., targets with substantially different binding sites). This can easily happen for protein families that contain a large and diverse set of proteins. The inclusion in the model of these less-relevant proteins can negatively impact the quality of the model learned for t_i. For example, in the case of the SVM-based approaches, the decision hyperplane may be unnecessarily minimizing the errors that are associated with the targets that are not relevant for t_i, whereas at the same time increasing the errors associated with t_i itself or other relevant targets. Third, for the cases in which a specific target t_i shares key characteristics related to ligand binding and recognition with proteins of other families, the intra-family focus of the chemogenomics-based approaches fails to take advantage of the relevant activity information provided by proteins in other families, leading to lower quality SAR models.

The affinity-based approaches that are developed in this paper are designed to overcome all three of the above shortcomings. For each specific target t_i, these approaches identify a set of protein targets that are related to t_i and then utilize only the activity information from these tar-

gets while learning t_i's SAR model. In addition, by using appropriate target-to-target similarity functions, these approaches can adapt to the characteristics of the individual protein targets and allow them to potentially select a subset of the proteins in t_i's family or proteins across different families. Finally, since these approaches do not rely on protein family membership, they can be used for proteins for which there is no activity information for any other family member.

The subsequent sections describe the different target-to-target similarity measures that we developed for identifying the set of related proteins and the different machine-learning methods that we developed for improving the quality of the target-specific SAR model by utilizing activity information from its related targets.

2.4 Identifying Related Targets

We developed two classes of target-to-target similarity functions that capture the similarity between the targets by taking into account two different types of information. The first takes into account the amino acid sequence of the targets whereas the second takes into account the similarity between their ligands.

2.4.1 Sequence-based Methods

Protein targets that have similar ligand binding sites in terms of their amino acid composition and their 3D structure show similar binding affinity towards a similar set of compounds.[32] Thus, a natural way of comparing two targets is to compare the sequences and structures of their binding sites. However, in many case the 3D structure of the proteins under consideration is not known (e.g., GPCRs), making it hard to accurately and reliably compare the ligand binding sites for all proteins. For this reason, we developed a target-to-target similarity function, referred to as $\mathcal{K}_t^{Seq}$, that measures the similarity between two protein targets by taking into account their entire amino acid sequences. Specifically, $\mathcal{K}_t^{Seq}(t_i, t_j)$ is computed as the optimal local alignment score[33] between t_i's and t_j's PSI-BLAST derived sequence profiles[34] and the PICASSO[35] profile-based scoring scheme. This profile-based alignment method combined with the PICASSO scoring scheme has been shown to better capture the evolutionary conserved sequence conservation signals between the proteins.[35,36]

2.4.2 Ligand-based Methods

The similarity between two targets can also be indirectly determined by considering their ligands. If two targets t_i and t_j have similar sets of ligands, then most likely their corresponding ligand binding sites share certain common characteristics. As a result, the similarity between their sets of ligands can be an implicit measure of the similarity of their binding sites. Motivated by this, we developed two approaches for determining the target-to-target similarity that take into account the similarity between their ligands. The first, referred to as $\mathcal{K}_t^{Aligs}$, measures the pairwise similarity of two targets t_i and t_j as the average pairwise similarity between their ligands. That is,

$$\mathcal{K}_t^{Aligs}(t_i, t_j) = \frac{\sum_{c_x \in C_i^+} \sum_{c_y \in C_j^+} \mathrm{sim}_c(c_x, c_y)}{|C_i^+||C_j^+|}. \tag{4}$$

The second, referred to as $\mathcal{K}_t^{kligs}$, measures the pairwise similarity of two targets t_i and t_j by considering only the average pairwise similarity of the k-nearest neighbors of each ligand to the other target's ligands. Specifically, $\mathcal{K}_t^{kligs}(t_i, t_j)$ is given by

$$\begin{aligned} \mathcal{K}_t^{kligs}(t_i, t_j) = & \frac{1}{k|C_i^+|} \sum_{c_x \in C_i^+} \sum_{c_y \in \mathrm{nbrs}_k(c_x, C_j^+)} \mathrm{sim}_c(c_x, c_y) \\ & + \\ & \frac{1}{k|C_j^+|} \sum_{c_x \in C_j^+} \sum_{c_y \in \mathrm{nbrs}_k(c_x, C_i^+)} \mathrm{sim}_c(c_x, c_y). \end{aligned} \tag{5}$$

The design of $\mathcal{K}_t^{kligs}$ was motivated by the fact that targets may contain ligands that come from multiple (and potentially different) scaffolds. As a result, the $\mathcal{K}_t^{Aligs}$ function, will unnecessarily penalize a pair of protein targets, each containing ligands derived from different scaffolds, even when the sets of scaffolds in each target are similar.

2.5 Affinity-based SAR Models using Semi-Supervised Learning

The first method that we developed for building an affinity-based SAR model for a specific target utilizes approaches based on semi-supervised learning.[18] The main idea of semi-supervised learning methods is to take advantage of the unlabeled instances during training in order to modify or re-prioritize the hypotheses obtained from the labeled instances.[37] This is usually done using a two-step process. In the first step, labels are assigned to the unlabeled instances. In the second step, a model is learned using both the original and the newly labeled instances.

Within the context of learning an affinity-based SAR model for a specific target t_i, the semi-supervised learning approach that we developed considers as unlabeled only those compounds that are ligands to at least one of the

related proteins and are neither positive nor negative instances for t_i. Specifically, if $R_i = \{t_{i_1}, t_{i_2}, \ldots, t_{i_m}\}$ are the m most similar target of t_i in $\mathbb{T}$ in non-increasing similarity order (i.e., the m related targets of t_i), then the set of compounds that are considered to be unlabeled is

$$U_i = \left(\bigcup_{1 \leq j \leq m} C_{i_j}^+ \right) \setminus (C_i^+ \cup C_i^-).$$

The motivation behind this definition is that the compounds in U_i corresponds to a biologically relevant subset of the chemical space as it contains compounds that have been experimentally determined to bind to a set of protein targets that are similar to the specific target t_i.

Details on how the labels are assigned to the compounds in U_i and how they are used to build better SAR models are provided in the next two sections.

2.5.1 Methods for Labeling Unlabeled Compounds

We developed two methods for labeling the compounds in U_i. The first method is based on a simple k-nearest neighbor scheme, whereas the second method employs an approach based on label propagation[38] that is used extensively for labeling unlabeled instances in semi-supervised learning.

In the k-nearest-neighbor-based method, referred to as $\text{LS}_{k\text{nn}}$, the compounds in U_i that belong in the k-nearest-neighbor list of at least one compound in C_i^+ (i.e., $\text{Nbrs}_k(C_i^+, U_i)$) are labeled as positives and the remaining compounds are labeled as negatives. This is motivated by the fact that compounds that are structurally similar tend to share the same biological activity.[39] As a result, those compounds in U_i that are similar to t_i's own ligands have a high probability of being active for t_i (i.e., be positive), whereas compounds that are dissimilar to t_i's ligands have a high probability of being inactive (i.e., be negative). Note that $\text{LS}_{k\text{nn}}$, is similar in spirit to the cluster kernel,[40] which assumes that unlabeled data within the neighborhood of the labeled data should be used with the same labels.

In the label propagation-based method, referred to as LS_{LP}, the labels of the compounds in U_i are determined by first constructing a weighted k-nearest neighbor compound-to-compound similarity graph involving both labeled and unlabeled compounds and then using an iterative procedure to propagate the labels from the labeled to the unlabeled nodes in this graph. Specifically, the graph contains a positively labeled node for each compound in C_i^+, a negatively labeled node for each compound in C_i^-, and an unlabeled node for each compound in U_i.

Program 1 LABEL_PROPAGATION($C_i^+ \cup C_i^-$, U_i)

```
1:  for each c_p, c_q ∈ C_i^+ ∪ C_i^- ∪ U_i do
2:          w_{p,q} = sim_c(c_p, c_q)
3:  end
4:  for each c_p, c_q ∈ C_i^+ ∪ C_i^- ∪ U_i do
5:          T_{p,q} = w_{p,q} / Σ_{k=1}^{|C|} w_{k,q}
6:  end
7:  for each c_p ∈ C_i^+ do
8:          L(p, 0) = 1
9:          L(p, 1) = 0
10: end
11: for each c_q ∈ C_i^- do
12:         L(p, 0) = 0
13:         L(p, 1) = 1
15: end
16: for each c_p ∈ U_i do
17:         L(q, 0) = L(q, 1) = 0.5
18: end
19: while(!L converges) do
20:         L ← TL
21:         Row-normalize L
22: end
23: for each c_p ∈ U_i do
24:         if L(p, 0) > L(p, 1)
25:             then  L(p, 0) = 1
26:                   L(p, 1) = 0
27:             else
28:                   L(p, 0) = 0
29:                   L(p, 1) = 1
30:         end
31: end
32: return L
```

The pseudo-code for the label propagation algorithm is shown in Algorithm 1. With $n = |C_i^+ \cup C_i^- \cup U_i|$, T is a $n \times n$ transition matrix, L is a $n \times 2$ label matrix, and $w_{p,q}$ is the weight assigned to the edge (p, q) that corresponds to the similarity between compounds p and q. The algorithm initially starts by computing the transition matrix (lines 1–6), initializing the labels of the nodes corresponding to the compounds in C_i^+ and C_i^- (lines 7–15), and assigning a weight of 0.5 to the labels for the rest of the nodes (lines 16–18). Then it proceeds to iteratively propagate the labels (lines 19–22) until convergence.[38] Finally, the labels of the nodes in U_i are determined as the maximum weight label (lines 23–31).

2.5.2 Building SAR Models Using the Newly Labeled Compounds

The labeling methods described in the previous section will assign a label (either positive or negative) to all the compounds in U_i. However, since the nature of the models

that we learn rely only on positively labeled instances (the negative instances are obtained by randomly sampling the unlabeled instances), we use only the positive subset of the newly labeled instances, denoted by H_i^+, as additional labeled instances to learn a SAR model for target t_i.

Specifically, we developed two different approaches for incorporating the newly labeled compounds into t_i's SAR model. The first approach, treats the original (C_i^+) and the newly labeled (H_i^+) positive instances equally, whereas the second approach, controls the influence that H_i^+'s compounds can have on the final model, by assigning a different misclassification cost to the compounds in C_i^+ and H_i^+. This differential weighting is done by using a variable α ($0 \leq \alpha \leq 1$) that controls the relative importance of C_i^+'s compounds over those in H_i^+ and then assigning a weight w_k to each compound c_k such that

$$w_k = \begin{cases} \alpha\left(1 + \frac{|H_i^+|}{|C_i^+|}\right) & \text{if } c_k \in C_i^+, \\ (1-\alpha)\left(1 + \frac{|C_i^+|}{|H_i^+|}\right) & \text{if } c_k \in H_i^+. \end{cases} \tag{6}$$

As a result, the compounds in C_i^+ will account for $\alpha\%$ of the overall weight of the positive instances, whereas the compounds in H_i^+ will account for the rest. Note that when $\alpha = |C_i^+|/|C_i^+ \cup H_i^+|$, this approach assigns a weight of one to all the compounds in C_i^+ and H_i^+, at which point it becomes identical to the first approach. We will denote these two approaches as CWS_{none} and CWS_α, respectively.

In addition, we also extended the CWS_α weighting scheme to take into account the similarity between t_i and its m related targets. The motivation behind this weighting scheme is to increase the importance of the compounds obtained from the most similar targets over the targets that are less similar. We used two methods for providing such weighting. The first, referred to as $\text{CWS}_\alpha^{\text{sim}}$, assigns a weight to compound $c_{l,j} \in H_i^+$, which was originally active against target t_{i_j} (i.e., $c_{l,j} \in C_{i_j}^+$), that is linearly proportional to the similarity between targets t_i and t_{i_j}. The second, referred to as $\text{CWS}_\alpha^{\text{exp}}$, assigns a weight to $c_{l,j}$ that decays exponentially with j (i.e., the rank of its target t_{i_j} in the list of m most similar targets of t_i). Note that when a compound in H_i^+ is active against more than one of the m most similar targets, it is only considered for its most similar target.

The precise weights assigned to the different compounds in conjunction with the differential weighting scheme of Equation 6 are as follows. For the $\text{CWS}_\alpha^{\text{sim}}$, the weight $w_{l,j}$ assigned to compound $c_{l,j}$ is given by

$$w_{l,j} = \frac{(1-\alpha)(|C_i^+| + |H_i^+|)}{\sum_{c_{r,q} \in H_i^+} \text{sim}_t(t_i, t_{i_q})} \text{sim}_t(t_i, t_{i_j}), \tag{7}$$

where $\text{sim}_t(t_i, t_{i_j})$ is the target-to-target similarity calculated from $\mathcal{K}_t^{Seq}$, $\mathcal{K}_t^{Aligs}$ or $\mathcal{K}_t^{kligs}$. For the $\text{CWS}_\alpha^{\text{exp}}$, the weight is given by

$$w_{l,j} = \frac{(1-\alpha)(|C_i^+| + |H_i^+|)}{\sum_{c_{r,q} \in H_i^+} 2^{-q}} 2^{-j}. \tag{8}$$

2.6 Affinity-based SAR Models using Multi-Task Learning

The second class of methods that we developed for building affinity-based SAR models for a specific target is based on multi-task learning.[19–22] Multi-task learning is a transfer learning mechanism designed to improve the generalization performance of a given model by leveraging the domain-specific information contained in the training signals of related tasks. In multi-task learning the model for a task (i.e., class) is learned in parallel with that of other related tasks using a shared representation so as to exploit dependencies between the tasks during learning. In recent years, various studies have reported promising results with the use of multi-task learning for various problems in cheminformatics.[12,17,41–44]

Motivated by the success of these methods, we developed a multi-task learning-based approach that leverages the activity information of the related targets. In this approach, the model for the specific target (t_i) is learned simultaneously with the models of its m related targets ($R_i = \{t_{i_1}, t_{i_2}, \ldots, t_{i_m}\}$) and the dependencies between the different targets and their ligands are captured via the use of target- and compound-specific kernel functions during SVM learning.

The input to this approach is a set of target-compound tuples (t_q, c_j) for each $t_q \in \{t_i\} \cup R_i$. For each target in $\{t_i\} \cup R_i$, tuples corresponding to target-ligand pairs (i.e., $c_j \in C_q^+$) are considered to be positive instances, whereas tuples corresponding to the non-binding compounds (i.e., $c_j \in C_q^-$) are considered to be negative instances. These tuples are used to train an SVM model $f()$ that learns how to separate the positive from the negative tuples. A SAR model for target t_i is then derived from $f()$ by computing $f((t_i, c))$ for each compound c whose activity against target t_i needs to be predicted.

Following the approach used by previously developed SVM-based approaches for learning multi-task mod-

els,[12,21,44] the dependencies between the different targets and compounds are coupled using a fusion-kernel based approach.[45] In this approach, the kernel function $\mathcal{K}_{mt}$ defined on the input target-compound tuples is given by

$$\mathcal{K}_{mt}((t_i, c_j), (t_{i'}, c_{j'})) = \beta\mathcal{K}_t(t_i, t_{i'}) + (1-\beta)\mathcal{K}_c(c_i, c_{i'}),$$

where $\mathcal{K}_t$ and $\mathcal{K}_c$ are kernel functions defined on the targets and the compounds, respectively, and β ($0 \leq \beta \leq 1$) is a parameter that weights the relative importance of the two components during training. The optimal value of β can be determined either during the learning phase[45–47] or empirically by performing a grid search over a set of values for these two parameters.[48] Note that the above formulation, by using a kernel function that combines both a target- and a compound-based component, allows SVM to capture relations between similar targets and their compounds and as such transfer knowledge across the different tasks during learning.

In order to use the above formulation, suitable target- and compound-based kernel functions need to be defined. For the target-based kernel function, we used the target-to-target similarity function (Section 2.4) that was used to identify the set of related proteins R_i. For example, if the set of related targets were identified using the $\mathcal{K}_t^{kligs}$ similarity function, then the same function was used as the target-based kernel. For the compound-based kernel function, we used the Tanimoto coefficient (Equation 1) as the kernel function for the compounds ($\mathcal{K}_c$) as it has been shown to produce good results for building SVM-based SAR models.

Note that a problem with the definitions of the $\mathcal{K}_t^{Seq}$ and $\mathcal{K}_t^{kligs}$ target-to-target similarity functions is that they lead to Gram-matrices that are symmetric but not necessarily positive semi-definite. For this reason they do not represent valid kernels and as such cannot be used directly for learning SVM models. To overcome this problem we use the approach described in Saigo *et al*[49] that converts a symmetric matrix into positive semi-definite by subtracting from the diagonal of the matrix its smallest negative eigenvalue. For the rest of the discussion, we will assume that this transformation has been applied to the $\mathcal{K}_t^{Seq}$ and $\mathcal{K}_t^{kligs}$ functions.

2.7 Affinity-based SAR Models using Multi-Ranking

Finally, motivated by classification approaches that determine the class of an unknown instance by combining predictions of a set of different classifiers, known as classification ensembles,[23–25] we developed an alternate method to improve the quality of a specific target's SAR model by taking advantage of the activity information of its m related targets $R_i = \{t_{i_1}, t_{i_2}, \ldots, t_{i_m}\}$. The main idea of this approach, referred to as *multi-ranking*, is to learn $m+1$ different SAR models, one for t_i and one for each target in R_i, use each of these models to compute a prediction score for an unknown compound c_i, and then determine the overall prediction score for c_i with respect to target t_i, by combining the $m+1$ individual predictions. The rationale behind this approach is that the SAR models of t_i's most similar targets should be able to detect the compounds that are active against t_i, and as such they can be used to re-enforce the predictions obtained from t_i's own SAR model.

In the multi-ranking methods, each of the $m+1$ SAR models are learned using the SVM-based framework described in Section 2.2. Specifically, for a target $t_j \in \{t_i\} \cup R_i$, its SAR model is learning using its active compounds C_j^+ as the positive instances and C_j^- as the negative instances. The $m+1$ individual prediction scores are combined by taking into account two factors. First, the relative importance of t_i's own prediction over that of its m related targets, and second, the relative importance of the m most similar targets among themselves. These two factors are similar to those discussed earlier in the context of semi-supervised learning for assigning different weights to the newly labeled compounds (Section 2.5.2).

We developed three different schemes for accounting for these two factors. Let s'_i be the prediction score for compound c_l obtained by t_i's SAR model, and let s'_{i_j} be the prediction scores obtained from the SAR models of t_i's m most similar targets. Then the overall prediction score s_i for the three schemes is given by

$$s_i = \alpha s'_i + \frac{1-\alpha}{m} \sum_{1 \leq j \leq m} s'_{i_j}, \tag{9}$$

$$s_i = \alpha s'_i + \frac{1-\alpha}{\sum_{1 \leq j \leq m} \text{sim}_t(t_i, t_{i_j})} \sum_{1 \leq j \leq m} \text{sim}_t(t_i, t_{i_j}) s'_{i_j}, \tag{10}$$

$$s_i = \alpha s'_i + \frac{1-\alpha}{\sum_{1 \leq j \leq m} 2^{-j}} \sum_{1 \leq j \leq m} 2^{-j} s'_{i_j}, \tag{11}$$

where α ($0 \leq \alpha \leq 1$) is a parameter that controls the relative importance of t_i's own prediction over the predictions obtained from the other targets to the overall prediction score. The predictions from the other targets are incorporated by the second term of the above equations. Equation 9 simply uses the average prediction score, whereas Equations 10 and 11 are based on the $\text{CWS}_\alpha^{\text{sim}}$ and $\text{CWS}_\alpha^{\text{exp}}$ schemes (Section 2.5.2), respectively. We refer to these three prediction combination

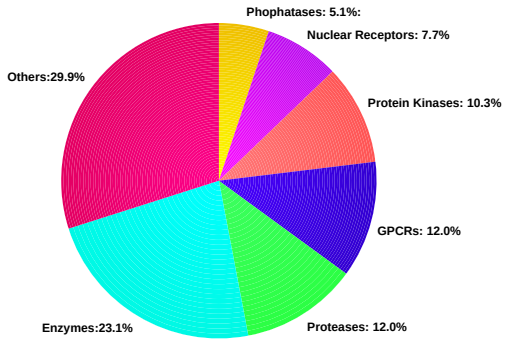

Figure 1: Distribution of protein targets

schemes as $\text{MWS}_\alpha^{\text{eql}}$, $\text{MWS}_\alpha^{\text{sim}}$ and $\text{MWS}_\alpha^{\text{exp}}$, respectively.

3 Materials

3.1 Datasets

We evaluated the performance of the affinity-based SAR models using a set of 146 protein targets and their ligands that were derived from various target-specific dose-response confirmatory screening assays. These screens were performed by NIH's Molecular Libraries Probe Production Centers Network (MLPCN) and are available in PubChem.[50] For each protein target its set of active compounds was determined using the activity assessment provided by the screening centers. Compounds that showed different activity signals in different assays against the same target were filtered out. For each of the protein targets, a baseline SAR model was learned and its performance was assessed using a five-fold cross validation. Since the goal of this work is to improve the performance of SAR models, we eliminated the set of targets for which their baseline SAR models achieved an ROC score greater of 0.80 (i.e., targets for which good models can be built by existing methods). This filtering resulted in 117 targets, 15,833 ligands, 16,088 target-ligand activity pairs (compounds can show activity against multiple protein targets), and an average of 138 active compounds per target. The distribution of the 117 protein targets in terms of their biological activity is shown in Figure 1.

3.2 Chemical Compound Descriptor

The chemical compounds were represented using the topological descriptors based on graph fragments (GF).[51] The GF descriptors correspond to all connected subgraphs up to a user-specified length that exist in a compound library. Comparisons against other popular topological descriptors (extended connectivity fingerprints, Maccs keys (MDL Information Systems Inc.), and frequent subgraph descriptors) have shown that the GF descriptors lead to a chemical compound representation that captures its structural characteristics effectively. As a result, its performance is either better than or comparable to that achieved by currently used descriptors for the tasks of building SVM-based SAR models and similarity searching. The GF descriptors were generated using the AFGEN[52] program and contained all the graph fragments of size four to seven bonds.

3.3 Support Vector Machines

We used the publicly available support vector machine tool SVM^{light} [53] that implements an efficient soft margin optimization algorithm. In all of our experiments, we used the default parameters for solving the quadratic programming problem and the default regularization parameter C that controls the margin width.

3.4 Evaluation Methodology & Metrics

The performance of the different methods was evaluated using a five-fold cross validation framework. For each target t_i, its set of positive C_i^+ and negative C_i^- compounds were split into five equal-size parts (folds). The compounds in each subset of four folds were used to train a model, which was then used to predict the compounds of the left-out fold.

The quality of the SAR models was measured using the ROC score,[54] which is the normalized area under the curve that plots the true positives against the false positives for different thresholds for classification (receiver operating characteristic curve). Since a five-fold cross validation was used, the computed ROC scores correspond to the average of the five folds. During the experimental evaluation we primarily report the average ROC improvements achieved by a method over the baseline models across the 117 protein targets. We used the geometric mean to compute these average improvements as it is better suited for averaging ratios.

4 Results

We performed a comprehensive study of the various parameters of the affinity-based methods described in Section 2 in order to assess the extent to which they lead to

SAR model improvements. In the rest of this section we present and summarize the key results from this study. All comparisons are done against the performance achieved by the baseline SAR models (Section 2.2). The results being presented correspond to some of the best performing combinations of the various parameters for the different schemes. The complete set of results are available as part of the supplementary material[1].

4.1 Performance of the Methods Based on Semi-Supervised Learning

Table 1 shows the average improvements achieved by the affinity-based semi-supervised learning methods over the baseline methods over the entire set of targets in our dataset. These results show that for certain parameter combinations the affinity-based semi-supervised learning approaches can achieve consistent improvements over the baseline model. The best performance achieved by the affinity-based semi-supervised learning approach is an average improvement of 1.8% ($\mathcal{K}_t^{Aligs}$ with LS_{LP}).

Comparing the performance achieved by the three target-to-target similarity measures we see that $\mathcal{K}_t^{Aligs}$ achieves consistently better results, whereas the results achieved by $\mathcal{K}_t^{Seq}$ are consistently the worst. The performance of $\mathcal{K}_t^{kligs}$ is between these two. These results suggest that the ligand-based similarity measures can better identify the proteins whose binding sites have similar characteristics than those based on sequence alignment scores. This is not surprising as the ligand-based similarity measures allow for the indirect comparison of the proteins binding sites, whereas the alignment-based methods may fail to compare the actual binding sites. One reason for the performance difference between $\mathcal{K}_t^{Aligs}$ and $\mathcal{K}_t^{kligs}$ is due to the differences in the number of unlabeled instances that exist in the sets of related targets identified by these two methods. The set of related targets identified by $\mathcal{K}_t^{kligs}$ results in a larger number of unlabeled instances (second column of Table 1) than the corresponding set for $\mathcal{K}_t^{Aligs}$. As a result, the number of positive labeled instances is larger for $\mathcal{K}_t^{kligs}$ than $\mathcal{K}_t^{Aligs}$ (columns labeled as $|H_i|$), which creates more diverse training sets that do not lead to good models. This difference between $\mathcal{K}_t^{Aligs}$ and $\mathcal{K}_t^{kligs}$ occurs because the former selects the related targets by taking into account all the ligands of the selected targets, whereas the latter looks at only the union of the k most similar ligands. As a result, ceteris paribus, targets with more ligands will be selected since they have a higher probability of containing a subset of compounds that are similar to the ligands of the target under consideration.

[1] http://www-users.cs.umn.edu/~xning/supplementary/

Comparing the performance of the two labeling schemes ($\text{LS}_{k\text{nn}}$ and LS_{LP}) we see that $\text{LS}_{k\text{nn}}$ tends to label as positive a smaller fraction of the unlabeled compound than LS_{LP}. Depending on the target-to-target similarity method being used, this can either lead to better or worse results. In the case of $\mathcal{K}_t^{kligs}$, for which the total number of unlabeled instances is large, the performance achieved by LS_{LP} is worse than that of $\text{LS}_{k\text{nn}}$, as it ends up labeling too many instances as positive. On the other hand, when $\mathcal{K}_t^{Aligs}$ is used, for which the total number of unlabeled instances is small, LS_{LP} performs better than $\text{LS}_{k\text{nn}}$. However, when the number of compounds that are being labeled by both schemes is approximately the same (e.g., $\mathcal{K}_t^{Aligs}$ and $m = 1, 3, 5$), the LS_{LP} achieves better results, suggesting that it does a better job in labeling the unlabeled compounds.

Comparing the performance of the different compound weighting schemes (CWS_{none} and $\text{CWS}_{\alpha}^{\text{exp}}$) we see that as the number of unlabeled compounds that is labeled as positive increases, $\text{CWS}_{\alpha}^{\text{exp}}$ does better than CWS_{none}. This is because under these conditions $\text{CWS}_{\alpha}^{\text{exp}}$, by decreasing the mis-classification weight of each newly labeled compound, reduces the overall influence that these compounds in the learned model. Also, not surprisingly, $\text{CWS}_{\alpha}^{\text{exp}}$'s performance improves when more weight is given to the original set of positive instances (i.e., the known ligands of each target) than the positive instances obtained as a result of the semi-supervised learning method (i.e., putative ligands).

Finally, comparing the performance of the schemes as the number m of related targets changes we see that, in general, their performance tends to initially improve as m increases and then it starts to degrade. Depending on the specific set of parameters, the best performance is usually achieved when 3–5 related targets are used. However, the methods based on $\mathcal{K}_t^{Seq}$ exhibit different performance characteristics as their performance consistently decreases as m increases.

4.2 Performance of the Methods Based on Multi-Task Learning

The average improvements achieved by the affinity-based multi-task learning methods over the baseline models are shown in Table 2. These results show that the ROC scores achieved by these models are usually higher than those achieved by the baseline model. Both the $\mathcal{K}_t^{Aligs}$ and $\mathcal{K}_t^{kligs}$ kernel functions achieve substantial improvements that range between 2.9% and 7.2%. Moreover, even the $\mathcal{K}_t^{Seq}$ kernel function, which in the context of semi-supervised learning (Table 1) always resulted in lower ROC scores

Table 1: Performance improvements of affinity-based semi-supervised learning methods.

			LS_{knn}						LS_{LP}					
	m	$\lvert U_i\rvert$	$\lvert H_i^+\rvert$	CWS_{none}	CWS_α^{exp}				$\lvert H_i^+\rvert$	CWS_{none}	CWS_α^{exp}			
					0.2	0.5	0.8	0.9			0.2	0.5	0.8	0.9
$\mathcal{K}_t^{Seq}$	1	86	49	-1.7%	-3.8%	-2.0%	-1.8%	-0.9%	68	-2.1%	-3.5%	-2.3%	-2.5%	-2.0%
	3	274	113	-2.1%	-3.0%	-2.2%	-1.3%	-1.0%	203	-5.8%	-6.5%	-5.9%	-5.8%	-4.1%
	5	449	146	-2.8%	-3.7%	-2.9%	-2.1%	-1.3%	367	-7.5%	-7.8%	-7.6%	-7.0%	-6.2%
	7	594	167	-2.9%	-2.6%	-1.7%	-2.2%	-1.6%	512	-7.3%	-7.3%	-7.1%	-7.3%	-6.6%
	9	752	182	-3.8%	-3.0%	-2.5%	-2.3%	-1.9%	621	-8.3%	-8.3%	-8.0%	-7.5%	-6.8%
$\mathcal{K}_t^{Aligs}$	1	41	26	1.1%	-0.4%	0.9%	0.8%	0.9%	28	**1.8%**	0.7%	1.7%	1.3%	0.8%
	3	122	70	0.7%	-1.0%	0.5%	**1.2%**	0.9%	78	1.3%	0.2%	1.5%	1.5%	**1.8%**
	5	216	106	-0.5%	-1.5%	-0.2%	0.6%	0.8%	122	0.8%	0.4%	0.8%	0.9%	1.5%
	7	317	134	-0.7%	-1.2%	-0.5%	0.4%	0.5%	243	-0.5%	-0.5%	-0.4%	0.6%	0.9%
	9	432	157	-1.1%	-1.2%	-0.4%	-0.5%	-0.1%	324	-1.3%	-1.0%	-0.9%	0.0%	0.8%
$\mathcal{K}_t^{kligs}$	1	114	61	0.7%	-0.4%	0.8%	0.5%	0.8%	89	-0.3%	-0.6%	0.0%	-0.9%	-0.8%
	3	364	135	-0.4%	-0.9%	-0.5%	-0.2%	0.2%	302	-0.5%	-0.5%	-0.5%	-0.5%	-0.4%
	5	625	179	-1.0%	-1.0%	-0.8%	-0.5%	0.6%	543	-1.5%	-1.5%	-1.3%	-1.2%	-1.1%
	7	894	208	-1.4%	-1.5%	-1.4%	-0.6%	-0.4%	703	-1.7%	-1.7%	-1.6%	-1.9%	-1.9%
	9	1181	229	-1.8%	-1.7%	-1.3%	-1.2%	-1.2%	945	-2.6%	-2.3%	-2.2%	-2.0%	-2.2%

In this table, m is the number of related targets, $|U_i|$ is the total number of unlabeled compounds, $|H_i^+|$ is the number of unlabeled compounds that were labeled as positive by the two labeling schemes (LS_{knn} and LS_{LP}). The columns labeled 0.2, 0.4, 0.8, and 0.9 correspond to the value of the α parameter for CWS_α^{exp}. The LS_{LP} was applied on the 5-nearest-neighbor graph of the labeled and unlabeled compounds. The $\mathcal{K}_t^{kligs}$ target-to-target similarity used $k = 5$. Bold-faced numbers indicate the best performing scheme under a certain combination of target-to-target similarity function and labeling scheme. Underlined numbers represent schemes with positive improvements.

than the baseline model, is able to achieve an improvement of 2.2% for $m = 1$ and $\beta = 0.1$.

Comparing the three kernel functions, we see that out of the 20 cases shown in Table 2, $\mathcal{K}_t^{kligs}$ achieves the best performance in 14 cases, $\mathcal{K}_t^{Aligs}$ in 6, whereas $\mathcal{K}_t^{Seq}$ never outperforms the other methods. The best overall performance is achieved by $\mathcal{K}_t^{kligs}$, which is a 7.2% improvement over the baseline model. The relatively poor performance of $\mathcal{K}_t^{Seq}$ over the ligand-based kernel functions is consistent with the earlier results involving semi-supervised learning and further re-enforces the fact that it is not well-suited for identifying appropriate targets for improving the accuracy of SAR models. However, in light of the results obtained by semi-supervised learning, the relative performance advantage of $\mathcal{K}_t^{kligs}$ over $\mathcal{K}_t^{Aligs}$ is somewhat surprising. This is due to the higher diversity among the targets identified by $\mathcal{K}_t^{kligs}$ and is further discussed later in Section 5. Comparing the performance of the ligand-based kernel functions as the number m of related targets increases, we observe that for $\mathcal{K}_t^{Aligs}$ and $\mathcal{K}_t^{kligs}$, the performance first improves and then degrades. The best performance is usually achieved when 3–5 related targets are used. However, for $\mathcal{K}_t^{Seq}$, as was the case with semi-supervised learning, the performance consistently decreases as m increases. Finally, comparing the performance of the two best-performing kernel functions as the value of β changes (Equation 2.6), we see that they exhibit distinctly different trends. The performance of $\mathcal{K}_t^{Aligs}$ remains largely unchanged as β ranges from 0.1 to 0.8, whereas the performance of $\mathcal{K}_t^{kligs}$ tends to markedly decrease for higher values of β. Thus, these results indicate that for $\mathcal{K}_t^{kligs}$, the best way to combine the target- and compound-based kernels in the fusion kernel formulation is by giving less weight to the target kernel and a higher-weight to the compound component.

4.3 Performance of Multi-Ranking

The average improvements achieved by the multi-ranking-based models over the baseline models are shown in Table 3. These results show that for a wide-range of parameter combinations, multi-ranking can achieve considerable improvements over the baseline models. The relative advantages of the three target-to-target similarity measures are consistent with the results obtained using the affinity-based multi-task learning method. $\mathcal{K}_t^{kligs}$ tends to perform the best, with some schemes achiev-

Table 3: Performance improvements of affinity-based multi-ranking methods.

	m	$\text{MWS}_\alpha^{\text{eql}}$				$\text{MWS}_\alpha^{\text{sim}}$				$\text{MWS}_\alpha^{\text{exp}}$			
		0.2	0.5	0.8	0.9	0.2	0.5	0.8	0.9	0.2	0.5	0.8	0.9
$\mathcal{K}_t^{Seq}$	1	-2.7%	1.7%	1.8%	1.3%	-2.7%	1.7%	1.8%	1.3%	-3.0%	1.7%	2.0%	1.3%
	3	-13.5%	0.7%	2.3%	1.5%	-11.8%	1.6%	2.5%	1.4%	-12.0%	0.2%	3.1%	1.9%
	5	-19.8%	-5.0%	2.1%	1.5%	-18.6%	-2.8%	2.1%	2.0%	-17.0%	-1.5%	3.0%	2.1%
	7	-20.2%	-9.8%	1.7%	1.2%	-19.4%	-6.3%	2.1%	1.7%	-18.0%	-2.1%	2.9%	2.1%
	9	-23.9%	-17.8%	2.1%	1.3%	-23.8%	-14.6%	2.7%	1.6%	-22.1%	-2.8%	2.8%	2.1%
$\mathcal{K}_t^{Aligs}$	1	1.3%	2.9%	2.4%	1.7%	1.3%	2.9%	2.4%	1.7%	1.0%	2.9%	2.2%	1.7%
	3	-6.6%	3.8%	2.5%	1.8%	-5.7%	4.2%	2.5%	1.9%	-4.9%	4.2%	3.1%	2.6%
	5	-12.1%	-0.4%	2.0%	1.8%	-11.7%	0.5%	2.1%	1.9%	-10.2%	2.7%	4.0%	2.9%
	7	-12.9%	-4.6%	3.0%	1.9%	-12.7%	-3.4%	3.1%	1.8%	-11.0%	2.1%	4.0%	3.0%
	9	-13.1%	-7.8%	4.6%	2.0%	-13.2%	-7.1%	4.4%	2.0%	-11.3%	2.4%	3.9%	3.0%
$\mathcal{K}_t^{kligs}$	1	0.8%	4.0%	4.2%	3.1%	0.8%	4.0%	4.2%	3.1%	1.0%	4.0%	4.0%	3.1%
	3	-5.1%	5.1%	5.3%	3.2%	-4.5%	5.8%	5.5%	3.2%	-3.3%	6.4%	6.0%	4.8%
	5	-10.2%	1.4%	6.3%	4.0%	-9.7%	2.4%	**6.4%**	4.1%	-6.2%	6.6%	**7.0%**	5.4%
	7	-13.6%	-5.4%	**6.5%**	4.4%	-13.4%	-4.6%	6.1%	4.4%	-11.0%	6.0%	**7.0%**	5.5%
	9	-16.2%	-10.6%	5.6%	3.4%	-16.1%	-9.8%	5.9%	3.6%	-14.2%	5.2%	**7.0%**	5.5%

In this table, m is the number of related targets. The columns labeled 0.2, 0.5, 0.8 and 0.9 correspond to the value of the α parameter for $\text{MWS}_\alpha^{\text{eql}}$, $\text{MWS}_\alpha^{\text{sim}}$ and $\text{MWS}_\alpha^{\text{exp}}$, respectively. The $\mathcal{K}_t^{kligs}$ target-to-target similarity used $k = 5$. Bold-faced numbers indicate the best performing scheme under a certain combination of target-to-target similarity function and prediction combination scheme. Underlined numbers represent schemes with positive improvements.

ing an average improvement of 7.0%, whereas $\mathcal{K}_t^{Seq}$ does relatively the worst, with its best performing parameter combination achieving only a 3.1% improvement. However, this 3.1% improvement is still substantially higher than the best performance achieved by any of the previous methods using this target-to-target similarity function or kernel function. Comparing the three prediction combination schemes $\text{MWS}_\alpha^{\text{eql}}$, $\text{MWS}_\alpha^{\text{sim}}$ and $\text{MWS}_\alpha^{\text{exp}}$, we see that on average $\text{MWS}_\alpha^{\text{exp}}$ performs the best, followed by $\text{MWS}_\alpha^{\text{sim}}$, and $\text{MWS}_\alpha^{\text{eql}}$ is the worst. This suggests that models from different targets do show different characteristics and function differently. Also, not surprising, the best performance is usually achieved when the original models contribute more to the overall prediction (i.e., $\alpha = 0.8$).

Comparing the performance of the multi-ranking approach as the number m of related targets increases, we observe that in general the performance initially improves and then it starts to degrade. The $\text{MWS}_\alpha^{\text{exp}}$ scheme is an exception, as in many cases its performance does not degrade. This is due to the exponential weighting on less similar targets which brings little impact on the combination of predictions. The best performance usually happens when 5–7 related targets are used. The degradation of performance associated with large m is because less similar models make less reliable predictions, and thus combining them will not introduce any benefits.

5 Discussion

The results presented in the previous section provide strong evidence that the affinity-based approaches can improve the quality of target-specific SAR models by utilizing activity information from related targets. When viewed together, these results point to the following trends. First, among the three affinity-based methods, multi-task learning and multi-ranking perform comparably and achieve the best results (Table 4 summarizes the best results of the three schemes). Second, the target-to-target similarity function that takes into account the entire sequence of the protein targets, does not perform as well as the ligand-based functions. This is due to the fact that the latter approaches indirectly account for the characteristics of the ligand binding sites, whereas the former does not. Third, the three affinity-based methods behave differently for the two ligand-based target-to-target similarity functions. Semi-supervised learning performs best for $\mathcal{K}_t^{Aligs}$, whereas the other two perform better for $\mathcal{K}_t^{kligs}$. As discussed in Section 4.1, $\mathcal{K}_t^{kligs}$ tends to select targets that have a large number of ligands. In the context of semi-supervised learning methods, this leads

Table 4: Summary of the performance improvements of the different affinity-based methods.

methods	target-to-target similarity	compound labeling	weighting scheme	weight	m	imprvmnt	ROC
semi-supervised learning	$\mathcal{K}_t^{Aligs}$	LS_{LP}	$\text{CWS}_{\alpha}^{\text{exp}}$	$\alpha = 0.9$	3	<u>1.8%</u>	0.66
multi-task learning	$\mathcal{K}_t^{kligs}$	-	-	$\beta = 0.2$	5	<u>**7.2%**</u>	0.70
multi-ranking	$\mathcal{K}_t^{kligs}$	-	$\text{MWS}_{\alpha}^{\text{exp}}$	$\alpha = 0.8$	5	<u>7.0%</u>	0.70

In this table, m is the number of related targets, *imprvmnt* is the best performance achieved by the affinity-based method under certain parameter combination in the corresponding row, and *ROC* is the average area under the ROC curve achieved by the corresponding scheme. The average ROC for baseline model was 0.65. Bold-faced numbers indicate the best performance over all affinity-based methods. Underlined numbers represent the schemes with positive improvements.

to a large number of unlabeled instances, which is the reason behind the lower performance of $\mathcal{K}_t^{kligs}$ over $\mathcal{K}_t^{Aligs}$. However, in the case of the methods based on multi-task learning and multi-ranking, this property of $\mathcal{K}_t^{kligs}$ actually leads to improved performance. This is because the targets selected by $\mathcal{K}_t^{kligs}$ tend to contain more diverse sets of compounds than those selected by $\mathcal{K}_t^{Aligs}$ (the average pairwise compound similarity of $\mathcal{K}_t^{Aligs}$'s five most similar targets was is 0.0138, whereas the corresponding similarity for $\mathcal{K}_t^{kligs}$ was only 0.0071) and consequently, there is a higher degree of diversity among the set of targets that are being selected by $\mathcal{K}_t^{kligs}$. This increase in diversity enables multi-task learning to exploit different areas of the chemical space during learning and enables multi-ranking to compute more robust predictions by averaging over less homogeneous models. Such increases in prediction heterogeneity are known to lead to performance improvements for ensemble-based methods.[23–25]

As discussed earlier, the quality of target-specific SAR models can be improved by using chemogenomics-based approaches that take into account the activity information from all the proteins in the same family. Within the context of the methods introduced in this paper, these chemogenomics-based approaches can be viewed as a special case of the affinity-based models in which all the proteins of the same family become the set of related proteins. Table 5 shows the performance gains achieved by the chemogenomics- and affinity-based approaches over the baseline models on the six protein families in our dataset that contain at least 4 members (i.e., the proteins in the "other" class of Figure 1 were not included). These results correspond to the parameter combinations that achieved the best performance for the different schemes. The machine learning methods that were used in the study are based on multi-task learning and multi-ranking, which outperform those based on semi-supervised learning. Results for three different schemes are provided, one using the chemogenomics approach (labeled "ChmGnmics") and two using the affinity-based approach (labeled "ABfamily" and "ABglobal"). The ABglobal method corresponds to the affinity-based schemes that were described in Section 2, whereas the ABfamily method corresponds to their variants in which the set of related targets were identified only from the same family. These results show that even though the chemogenomics-based approaches are able to improve the quality of the target-specific SAR models, these improvements are smaller than those obtained by the affinity-based approaches. Averaged over the 82 proteins in these six families, the affinity-based approaches achieve a 4.33% improvement over the chemogenomics-based approaches (best affinity-based scheme vs best chemogenomics-based scheme). In addition, comparing the performance achieved by the ABfamily variant of the affinity-based methods we see that they perform 0.9% better than the chemogenomics-based approaches and 3.3% worse than the actual affinity-based approaches (ABglobal). These results show that higher performance gains can be obtained by not utilizing the activity information from all the proteins in the family (ABfamily vs ChmGnmics) and that even further gains can be achieved by utilizing activity information from proteins of different families (ABglobal vs ABfamily).

To illustrate the cross-family nature of the affinity-based methods, Figure 2 shows the set of related proteins for the different proteins within and across the different families ($\mathcal{K}_t^{kligs}$ and $m = 3$). This figure shows that for

Table 5: Performance of chemogenomics- and affinity-based approaches relative to the baseline models.

family	scheme	Multi-Task Learning				Multi-Ranking				
		target-to-target similarity	β	m	imprvmnt	target-to-target similarity	weighting scheme	α	m	imprvmnt
Phosphatases	ChmGnmics	$\mathcal{K}_t^{Aligs}$	0.2	6	-5.2%	$\mathcal{K}_t^{Seq}$	$\text{MWS}_\alpha^{\text{sim}}$	0.9	6	0.2%
	ABfamily	$\mathcal{K}_t^{kligs}$	0.2	1	1.2%	$\mathcal{K}_t^{kligs}$	$\text{MWS}_\alpha^{\text{exp}}$	0.9	3	-2.3%
	ABglobal	$\mathcal{K}_t^{Aligs}$	0.2	1	**6.9%**	$\mathcal{K}_t^{Aligs}$	$\text{MWS}_\alpha^{\text{exp}}$	0.8	3	6.5%
Nuclear Receptors	ChmGnmics	$\mathcal{K}_t^{kligs}$	0.2	9	14.6%	$\mathcal{K}_t^{kligs}$	$\text{MWS}_\alpha^{\text{exp}}$	0.8	9	11.8%
	ABfamily	$\mathcal{K}_t^{kligs}$	0.2	3	14.0%	$\mathcal{K}_t^{kligs}$	$\text{MWS}_\alpha^{\text{exp}}$	0.5	3	13.3%
	ABglobal	$\mathcal{K}_t^{kligs}$	0.2	5	**18.1%**	$\mathcal{K}_t^{kligs}$	$\text{MWS}_\alpha^{\text{exp}}$	0.8	5	10.6%
Protein Kinases	ChmGnmics	$\mathcal{K}_t^{Aligs}$	0.5	12	4.3%	$\mathcal{K}_t^{Aligs}$	$\text{MWS}_\alpha^{\text{exp}}$	0.8	12	8.2%
	ABfamily	$\mathcal{K}_t^{kligs}$	0.2	5	11.3%	$\mathcal{K}_t^{kligs}$	$\text{MWS}_\alpha^{\text{sim}}$	0.8	7	9.8%
	ABglobal	$\mathcal{K}_t^{kligs}$	0.2	1	**15.3%**	$\mathcal{K}_t^{kligs}$	$\text{MWS}_\alpha^{\text{exp}}$	0.8	7	14.5%
GPCRs	ChmGnmics	$\mathcal{K}_t^{Aligs}$	0.2	14	-3.6%	$\mathcal{K}_t^{kligs}$	$\text{MWS}_\alpha^{\text{exp}}$	0.8	14	1.4%
	ABfamily	$\mathcal{K}_t^{Aligs}$	0.2	1	1.4%	$\mathcal{K}_t^{Aligs}$	$\text{MWS}_\alpha^{\text{sim}}$	0.8	7	2.6%
	ABglobal	$\mathcal{K}_t^{Aligs}$	0.2	3	**6.8%**	$\mathcal{K}_t^{kligs}$	$\text{MWS}_\alpha^{\text{exp}}$	0.8	1	3.2%
Proteases	ChmGnmics	$\mathcal{K}_t^{Aligs}$	0.2	14	0.8%	$\mathcal{K}_t^{kligs}$	$\text{MWS}_\alpha^{\text{exp}}$	0.8	14	5.1%
	ABfamily	$\mathcal{K}_t^{kligs}$	0.2	1	6.7%	$\mathcal{K}_t^{kligs}$	$\text{MWS}_\alpha^{\text{exp}}$	0.8	3	6.4%
	ABglobal	$\mathcal{K}_t^{kligs}$	0.2	5	**12.1%**	$\mathcal{K}_t^{kligs}$	$\text{MWS}_\alpha^{\text{exp}}$	0.8	7	9.3%
Enzymes	ChmGnmics	$\mathcal{K}_t^{Aligs}$	0.2	27	-6.6%	$\mathcal{K}_t^{kligs}$	$\text{MWS}_\alpha^{\text{exp}}$	0.8	27	1.3%
	ABfamily	$\mathcal{K}_t^{Aligs}$	0.2	1	0.4%	$\mathcal{K}_t^{kligs}$	$\text{MWS}_\alpha^{\text{exp}}$	0.8	7	1.3%
	ABglobal	$\mathcal{K}_t^{kligs}$	0.2	5	2.7%	$\mathcal{K}_t^{kligs}$	$\text{MWS}_\alpha^{\text{exp}}$	0.8	9	**2.9%**

In this table, β is the weight on target similarity for affinity-based multi-task learning method, α is the weight for $\text{MWS}_\alpha^{\text{eql}}$, $\text{MWS}_\alpha^{\text{sim}}$ and $\text{MWS}_\alpha^{\text{exp}}$, m is either the number of related targets (affinity-based approaches) or size of the protein family (chemogenomics-based approaches), *imprvmnt* is the performance of certain affinity-based scheme under corresponding combination of parameters for each protein family. *ChemGnmics* denotes the results obtained by the chemogenomics-based approach, *ABfamily* denotes the results obtained by the family-focused affinity-based approach, and *ABglobal* denotes the results obtained by the actual affinity-based approach. Bold-faced numbers indicate the best performing scheme for each protein family.

nearly all protein targets, a fair number of their related targets (66.5%) come from targets that belong to other families and include proteins that are substantially different from each other (e.g., kinases and GPCRs).

6 Conclusion

In this paper, we developed various machine learning methods to improve the quality of the SAR models for a given target by taking into account activity information from other targets. These methods include approaches based on semi-supervised learning that deliberately incorporate selected unlabeled compounds while training the SAR models, approaches based on multi-task learning that attempt to improve the quality of the SAR model by transferring information learned from different targets, and approaches based on multi-ranking that utilize the SAR models of different targets by relying on classifier ensembles. The comprehensive experimental evaluation of these methods on a large dataset of 117 protein targets have shown that substantial performance gains can be obtained as long as the set of targets from which activity information is utilized is properly selected. Among the methods developed, approaches based on multi-task learning and multi-ranking achieve the best overall performance, resulting in a 7.0%–7.2% average improvement over the performance achieved by the standard approaches for building SAR models that rely only on the activity information of the target under consideration. Moreover,

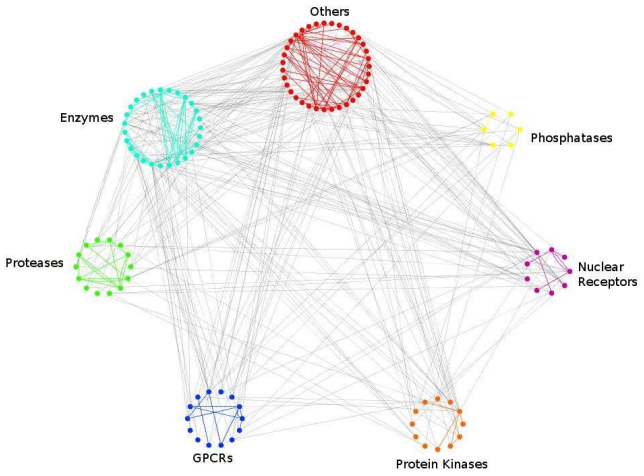

Figure 2: Connectivity pattern between the related proteins ($m = 3$) for the different families.

these methods by selecting the targets from which activity information will be utilized from the entire dataset, they outperform approaches based on chemogenomics that utilize activity information of protein targets that belong to the same family as that under consideration.

Acknowledgment

This work was supported by IIS-0431135, NIH RLM008713A, and by the Digital Technology Center at the University of Minnesota.

References

1. Hansch, C.; Maolney, P. P.; Fujita, T.; Muir, R. M. *Nature* **1962,** *194*, 178-180.
2. Hansch, C.; Muir, R. M.; Fujita, T.; Maloney, C. F.; M., S. *Journal of American Chemical Society* **1963,** *85*, 2817-1824.
3. Bravi, G.; Green, E. G. . D.; Hann, V.; Mike, M. Modelling Structure-Activity Relationship. In *Virtual Screening for Bioactive Molecules*, Vol. 10; Bohm, H.; Schneider, G., Eds.; Wiley-VCH: New York, NY, USA, 2000.
4. Agrafiotis, D.; Bandyopadhyay, D.; Wegner, J.; vanVlijmen, H. *Journal of Chemical Information and Modeling* **2007,** *47*, 1279-1293.
5. Warmuth, M. K.; Liao, J.; Ratsch, G.; Mathieson, M.; Putta, S.; Lemmen, C. *Journal of Chemical Information and Computer Sciences* **2003,** *43*, 667-673.
6. Frye., S. *Chemistry and Biology* **1999,** R3–R7.
7. Caron, P. R.; Mullican, M. D.; Mashal, R. D.; Wilson, K. P.; Su, M. S.; Murcko, M. A. *Curr Opin Chem Biol* **2001,** *5*, 464-70.
8. Klabunde, T. *Br J Pharmacol* **2007,** *152*, 5–7.
9. Bock, J.; Gough, D. *Journal of Chemical Information and Modeling* **2005,** *45*, 1402-1414.
10. Lapinsh, M.; Prusis, P.; Uhlen, S.; Wikberg, J. E. S. *Bioinformatics* **2005,** *21*, 4289-4296.
11. Lindström, A.; Pettersson, F.; Almqvist, F.; Berglund, A.; Kihlberg, J.; Linusson, A. *Journal of Chemical Information and Modeling* **2006,** *46*, 1154-1167.
12. Jacob, L.; Hoffmann, B.; Stoven, V.; Vert, J.-P. *BMC Bioinformatics* **2008,** *9*, 363.
13. Strömbergsson, H.; Daniluk, P.; Kryshtafovych, A.; Fidelis, K.; Wikberg, J. E. S.; Kleywegt, G. J.; Hvidsten, T. R. *Journal of Chemical Information and Modeling* **2008,** *48*, 2278-2288.
14. Deng, Z.; Chuaqui, C.; Singh, J. *J Med Chem* **2004,** *47*, 337–344.
15. Weill, N.; Rognan, D. *Journal of Chemical Information and Modeling* **2009,** *49*, 1049-1062.
16. Geppert, H.; Humrich, J.; Stumpfe, D.; Gartner, T.; Bajorath, J. *Journal of Chemical Information and Modeling* **2009,** *49*, 767-779.
17. Erhan, D.; L'Heureux, P.-J.; Yue, S. Y.; Bengio, Y. *Journal of Chemical Information and Modeling* **2006,** *46*, 626-635.
18. Chapelle, O.; Schölkopf, B.; Zien, A., Eds.; *Semi-Supervised Learning;* MIT Press: Cambridge, MA, 2006.
19. Thurn, S. Is learning the *n*-th thing any easier than learning the first?. In *Advances in Neural Information Processing Systems*; The MIT Press: NIPS Foundation, 1996.

Table 2: Performance improvements of the affinity-based multi-task learning methods.

	m	β			
		0.1	0.2	0.5	0.8
$\mathcal{K}_t^{Seq}$	1	2.2%	1.8%	0.8%	0.5%
	3	1.1%	0.8%	-0.8%	-1.8%
	5	-0.4%	-0.7%	-1.8%	-2.8%
	7	-0.5%	-1.0%	-1.8%	-3.1%
	9	-0.8%	-1.0%	-2.5%	-4.2%
$\mathcal{K}_t^{Aligs}$	1	3.3%	3.4%	3.2%	3.2%
	3	5.9%	5.9%	5.7%	5.8%
	5	5.4%	5.4%	5.3%	5.1%
	7	4.9%	5.0%	4.8%	4.4%
	9	4.9%	5.0%	4.8%	4.4%
$\mathcal{K}_t^{kligs}$	1	4.3%	3.7%	3.1%	2.9%
	3	7.0%	7.1%	4.9%	4.0%
	5	7.0%	**7.2%**	5.5%	4.1%
	7	6.4%	6.8%	5.3%	3.5%
	9	6.6%	6.9%	5.2%	3.4%

In this table, m is the number of related targets. The columns labeled 0.1, 0.2, 0.5 and 0.8 correspond to the value of the β parameter (i.e., weight on the target-based kernel). The $\mathcal{K}_t^{kligs}$ target-to-target similarity used $k = 5$. Bold-faced numbers indicates the best performance of affinity-based multi-task learning. Underlined numbers represent schemes with positive improvements.

20. Caruana, R. A. Multitask learning: A knowledge-based source of inductive bias. In *Proceedings of the Tenth International Conference on Machine Learning*; Morgan Kaufmann: MA, USA, 1993.
21. Evgeniou, T.; Micchelli, C. A.; Pontil, M. *Journal of Machine Learning Research* **2005,** *6*, 615–637.
22. Bonilla, E.; Agakov, F.; Williams, C. **2007,** .
23. Swanson, R.; Tsai, J. *J. Bacteriol.* **2003,** *185*, 3990-3993.
24. Kuncheva, L. I.; Whitaker, C. J. *Mach. Learn.* **2003,** *51*, 181–207.
25. Shipp, C. A.; Kuncheva, L. I. *Information Fusion* **2002,** *3*, 135 - 148.
26. P. Willett, J.; G.M.Downs, *Journal of Chemical Information and Computer Sciences* **1998,** *38*, 983–997.
27. Vapnik, V. *Statistical Learning Theory;* John Wiley: New York, 1998.
28. H., Y.; J., H.; K., C. PEBL: Positive example based learning for Web page classification using SVM. In *ACM KDD*; ACM: New York, NY, USA, 2002.
29. Liu, B.; Dai, Y.; Li, X.; Lee, W.; Yu, P. Building text classifiers using positive and unlabeled examples. In *Proceedings of the 3rd ICDM Conference*; IEEE Computer Society Press: Florida, USA, 2003.
30. Wang, C.; Ding, C.; Meraz, R. F.; Holbrook, S. R. *Bioinformatics* **2006,** *22*, 2590–2596.
31. Elkan, C.; Noto, K. Learning classifiers from only positive and unlabeled data. In *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*; ACM: New York, NY, USA, 2008.
32. Davies, E. K. *American Chemical Society* **1996,** *118*, 309-316.
33. Smith, T. F.; Waterman, M. S. *Journal of Molecular Biology* **1981,** *147*, 195–197.
34. Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. *Nucleic Acids Research* **1997,** *25*, 3389–3402.
35. Heger, A.; Holm, L. *Bioinformatics* **2001,** *17*, 272–279.
36. Rangwala, H.; Karypis, G. *Bioinformatics* **2005,** *21*, 4239-4247.
37. Zhu, X. "Semi-Supervised Learning Literature Survey", Technical Report, Computer Sciences, University of Wisconsin-Madison, 2005.
38. Zhu, X.; Ghahramani, Z. "Learning from labeled and unlabeled data with label propagation", Technical Report CMU-CALD-02-107, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, 2002.
39. Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. *Journal of Medicinal Chemistry* **2002,** *45*, 4350-4358.
40. Weston, J.; Elisseeff, A.; Zhou, D.; Leslie, C.; Noble, W. S. *PNAS USA* **2004,** *101*, 6559–6563.
41. Tetko, I. V.; Tanchuk, V. Y. *J Chem Inf Comput Sci* **2002,** *42*, 1136–1145.
42. Tetko, I. V.; Jaroszewicz, I.; Platts, J. A.; Kuduk-Jaworska, J. *J Inorg Biochem* **2008,** *102*, 1424–1437.
43. Varnek, A.; Gaudin, C.; Marcou, G.; Baskin, I.; Pandey, A. K.; Tetko, I. V. *J Chem Inf Model* **2009,** *49*, 133–144.
44. Jacob, L.; Hoffmann, B.; Stoven, V.; Vert, J.-P. *BMC Bioinformatics* **2008,** *9*, 363.
45. Lanckriet, G. R.; Deng, M.; Cristianini, N.; Jordan, M. I.; Noble, W. S. *Pac Symp Biocomput* **2004,** 300-11.
46. Sonnenburg, S.; Ratsch, G.; Schafer, C. *Proceedings of the 2005 Neural Information Processing Systems* **2005,** .
47. Tsang, I. W.; Kwok, J. T. . *IEEE Transactions on Neural Networks* **2006,** *17*, 48–58.
48. Rangwala, H.; Karypis, G. fRMSDAlign: Protein Sequence Alignment Using Predicted Local Structure Information for Pairs with Low Sequence Identity.. In *APBC*, Vol. 6; Brazma, A.; Miyano, S.; Akutsu, T., Eds.; Imperial College Press: Kyoto, Japan, 2008.
49. Saigo, H.; Vert, J.-P.; Ueda, N.; Akutsu, T. *Bioinformatics* **2004,** *20*, 1682-1689.

50. Austin, C. P.; Brady, L. S.; Insel, T. R.; Collins, F. S. *Science* **2004,** *306*, 1138–1139.

51. Wale, N.; Watson, I. A.; Karypis, G. *Knowledge and Information Systems* **2008,** *14*, 347–375.

52. Wale, N.; Karypis, G. "AFGEN", Technical Report, Department of Computer Science & Enigneering, University of Minnesota, 2007 www.cs.umn.edu/ karypis.

53. Joachims, T. *MIT press* **1999,** .

54. Fawcett, T. "ROC Graphs: Notes and Practical Considerations for Researchers", 2004.