

# Multi-Assay-Based Structure–Activity Relationship Models: Improving Structure–Activity Relationship Models by Incorporating Activity Information from Related Targets

Xia Ning,<sup>\*,†</sup> Huzefa Rangwala,<sup>‡</sup> and George Karypis<sup>†</sup>

Department of Computer Science and Computer Engineering, University of Minnesota, 4-192 EE/CS Building, 200 Union Street SE, Minneapolis, Minnesota 55455 and Department of Computer Science, George Mason University, 4400 University Drive MSN 4A5, Fairfax, Virginia 22030

Received May 19, 2009

Structure–activity relationship (SAR) models are used to inform and to guide the iterative optimization of chemical leads, and they 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 multi-assay 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 multi-assay-based methods achieve a receiver-operating characteristic 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 multi-assay-based methods outperform chemogenomics-based approaches by 4.33%.

## 1. INTRODUCTION

The pioneering work of Hansch et al.,<sup>1,2</sup> 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 SAR models.<sup>3,4</sup> These *in silico* models have become an essential tool for predicting the biological activity of a compound from its molecular structure and have 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.<sup>5</sup> In this approach, the target's experimentally determined ligands are used to build an initial support vector machine (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.<sup>6–8</sup> The key idea behind these approaches is to synergistically use information across a set of proteins that belong to the same family (e.g., G-protein-coupled receptors (GPCRs), kinases, etc.). 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 of all or a subset of the family members, better models can be expected. In these approaches, a model is trained using instances consisting of target ligand pairs from protein members of the same 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 by the features of the targets, the compounds, and the complexes that they utilize (e.g., physicochemical properties,<sup>9,10</sup> protein structure,<sup>11</sup> amino acid sequence,<sup>12</sup> binding site descriptors,<sup>13,14</sup> topological descriptors,<sup>9</sup> protein–ligand fingerprints,<sup>15</sup> etc.), the machine learning methods that they use for learning the models (e.g., support vector machines,<sup>12,16</sup> neural networks,<sup>17</sup> partial least-squares,<sup>10,11,14</sup> random forests,<sup>15</sup> multi-task learning,<sup>16</sup> etc.) and by how they represent target compound

\* Corresponding author. E-mail: xning@cs.umn.edu.

<sup>†</sup> University of Minnesota.

<sup>‡</sup> George Mason University.

pairs (e.g., concatenation of descriptor vectors,<sup>9</sup> tensor products,<sup>12</sup> kernel fusion,<sup>12</sup> 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 multi-assay 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 can 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 multi-assay-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,<sup>18</sup> multi-task learning,<sup>19–22</sup> and classifier ensembles.<sup>23–25</sup> The experimental evaluation of these methods on a set of 117 targets, extracted from PubChem, shows that for nearly all of them, the incorporation of activity information from other targets leads to quality improvements in the resulting multi-assay-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% receiver-operating characteristic (ROC) improvement. In addition, on a set of six protein families, the multi-assay-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  $T$ , and the union of active compounds over all targets by  $C$  (i.e.,  $C = \cup C_i^+$ ).

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.<sup>3</sup> 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.<sup>26</sup> 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 (multi-assay-based SAR model).

**2.2. Baseline SAR Models.** For each target  $t_i$ , we used support vector machines (SVM)<sup>27</sup> to build the baseline SAR model that relies only on its own activity information. Given a set of positive training instances  $I^+$  and a set of negative training instances  $I^-$ , SVM learns a classification function  $f(x)$  of the form

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

where  $\lambda_i^+$  and  $\lambda_i^-$  are non-negative weights that are computed during training by maximizing a quadratic objective function, and  $K$  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 nonincreasing 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 nonbinding 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<sup>28–31</sup> (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.<sup>4</sup> Recent work has shown that, under the assumption 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.<sup>31</sup>

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  $\mathcal{V} \setminus C_i^+$ . This allows for the creation of equal-size positive and negative training sets. Moreover, by using  $\mathcal{V} \setminus C_i^+$  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 multi-assay-based methods that are described in the subsequent sections.

**2.3. Multi-Assay-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 proteins in the same family. However, the fundamental step in these approaches, which is building a SAR model based on all or a subset of 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 that bind to substantially different sets of ligands). This can easily happen for protein families that contain a 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 intrafamily 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 multi-assay-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 targets 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 the 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 three-dimensional (3D) structure, show similar binding affinity toward a similar set of compounds.<sup>32</sup> Thus, a natural way of comparing two targets is to compare the sequences and structures of their binding sites. However, in many cases, 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  $K_i^{\text{seq}}$ , that measures the similarity between two protein targets by taking into account their entire amino acid sequences. Specifically,  $K_i^{\text{seq}}(t_i, t_j)$  is computed as the optimal local alignment score<sup>33</sup> between  $t_i$ 's and  $t_j$ 's PSI-BLAST derived sequence profiles<sup>34</sup> and the PICASSO<sup>35</sup> 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.<sup>35,36</sup>

**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  $K_i^{\text{aligs}}$ , measures the pairwise similarity of two targets  $t_i$  and  $t_j$  as the average pairwise similarity between their ligands. That is,

$$K_i^{\text{aligs}}(t_i, t_j) = \frac{\sum_{c_x \in C_i^+} \sum_{c_y \in C_j^+} \text{sim}_c(c_x, c_y)}{|C_i^+||C_j^+|} \quad (4)$$

The second, referred to as  $K_i^{\text{kligs}}$ , 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,  $K_i^{\text{kligs}}(t_i, t_j)$  is given by

$$K_i^{\text{kligs}}(t_i, t_j) = \frac{1}{k|C_i^+|} \sum_{c_x \in C_i^+} \sum_{c_y \in \text{nbrs}_k(c_x, C_j^+)} \text{sim}_c(c_x, c_y) + \frac{1}{k|C_j^+|} \sum_{c_x \in C_j^+} \sum_{c_y \in \text{nbrs}_k(c_x, C_i^+)} \text{sim}_c(c_x, c_y) \quad (5)$$

The design of  $K_i^{\text{kligs}}$  was motivated by the fact that targets may contain ligands that come from multiple (and potentially different) scaffolds. As a result, the  $K_i^{\text{aligs}}$  function will unnecessarily penalize a pair of protein targets, each contain-



ing ligands derived from different scaffolds, even when the sets of scaffolds in each target are similar.

**2.5. Multi-Assay-Based SAR Models using Semi-Supervised Learning.** The first method that we developed for building a multi-assay-based SAR model for a specific target utilizes approaches based on semi-supervised learning.<sup>18</sup> The main idea of semi-supervised learning methods is to take advantage of the unlabeled instances during training in order to modify or reprioritize the hypotheses obtained from the labeled instances.<sup>37</sup> 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, a multi-assay-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 that are neither positive nor negative instances for  $t_i$ . Specifically, if  $R_i = \{t_{i1}, t_{i2}, \dots, t_{im}\}$  is the  $m$  most similar target of  $t_i$  in  $T$  in nonincreasing similarity order (i.e., the  $m$  related targets of  $t_i$ ), then the set of compounds that are considered to be unlabeled is

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

The motivation behind this definition is that the compounds in  $U_i$  correspond 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<sup>38</sup> that is used extensively for labeling unlabeled instances in semi-supervised learning.

In the  $k$ -nearest-neighbor-based method (see Chart 1), referred to as  $LS_{knn}$ , the compounds in  $U_i$  that belong in the  $k$ -nearest-neighbor list of at least one compound in  $C_i^+$  (i.e.,  $Nbrs_k(C_i^+, U_i)$ ) 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.<sup>39</sup> 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., being positive), whereas compounds that are dissimilar to  $t_i$ 's ligands have a high probability of being inactive (i.e., being negative). Note that  $LS_{knn}$  is similar in spirit to the cluster kernel,<sup>40</sup> 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 by 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

Chart 1

**Program 1** LABEL-PROPAGATION( $C_i^+ \cup C_i^-, U_i$ )

---

```

1:  for each  $c_p, c_q \in C_i^+ \cup C_i^- \cup U_i$  do
2:       $w_{p,q} = \text{sim}_c(c_p, c_q)$ 
3:  end
4:  for each  $c_p, c_q \in C_i^+ \cup C_i^- \cup U_i$  do
5:       $T_{p,q} = \frac{w_{p,q}}{\sum_{k=1}^{|C_i^+|} w_{k,q}}$ 
6:  end
7:  for each  $c_p \in C_i^+$  do
8:       $L(p, 0) = 1$ 
9:       $L(p, 1) = 0$ 
10: end
11: for each  $c_q \in C_i^-$  do
12:      $L(p, 0) = 0$ 
13:      $L(p, 1) = 1$ 
14: end
15: for each  $c_p \in U_i$  do
16:      $L(q, 0) = L(q, 1) = 0.5$ 
17: end
18: while ( $L$  converges) do
19:      $L \leftarrow TL$ 
20:     Row-normalize  $L$ 
21: end
22: for each  $c_p \in U_i$  do
23:     if  $L(p, 0) > L(p, 1)$ 
24:         then  $L(p, 0) = 1$ 
25:          $L(p, 1) = 0$ 
26:     else
27:          $L(p, 0) = 0$ 
28:          $L(p, 1) = 1$ 
29:     end
30: end
31: return  $L$ 

```

---

negatively labeled node for each compound in  $C_i^-$ , and an unlabeled node for each compound in  $U_i$ .

The pseudocode for the label propagation algorithm is shown in Program 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.<sup>38</sup> 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  $\alpha$  ( $0 \leq \alpha \leq 1$ ) that controls the relative importance of  $C_i^+$ 's compounds over those in  $H_i^+$  and then by assigning a weight  $w_k$  to each compound  $c_k$  such that:

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

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 1 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  $CWS_{\text{none}}$  and  $CWS_{\alpha}$ , respectively.

In addition, we also extended the  $CWS_{\alpha}$  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  $CWS_{\alpha}^{\text{sim}}$ , 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  $CWS_{\alpha}^{\text{exp}}$ , 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  $CWS_{\alpha}^{\text{sim}}$ , the weight  $w_{l,j}$  assigned to compound  $c_{l,j}$  is given by

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

where  $\text{sim}_t(t_i, t_{i_j})$  is the target-to-target similarity calculated from  $K_t^{\text{seq}}$ ,  $K_t^{\text{align}}$ , or  $K_t^{\text{kligs}}$ . For the  $CWS_{\alpha}^{\text{exp}}$ , the weight is given by

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

**2.6. Multi-Assay-based SAR Models using Multi-Task Learning.** The second class of methods that we developed for building multi-assay-based SAR models for a specific target is based on multi-task learning.<sup>19–22</sup> 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.<sup>12,17,41–43</sup>

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 ( $R_i = \{t_{i_1}, t_{i_2}, \dots, t_{i_m}\}$ ), 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 nonbinding 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 models,<sup>12,21</sup> the dependencies between the different targets and compounds are coupled using a fusion-kernel based approach.<sup>44</sup> In this approach, the kernel function  $K_m$ , defined on the input target-compound tuples, is given by

$$K_m((t_i, c_j), (t_i, c_j)) = \beta K_t(t_i, t_i) + (1 - \beta) K_c(c_i, c_i)$$

where  $K_t$  and  $K_c$  are kernel functions defined on the targets and the compounds, respectively, and  $\beta$  ( $0 \leq \beta \leq 1$ ) is a parameter that weights the relative importance of the two components during training. The optimal value of  $\beta$  can be determined either during the learning phase<sup>44–46</sup> or empirically by performing a grid search over a set of values for these two parameters.<sup>47</sup> Note that the above formulation, by using a kernel function that combines both target- and compound-based components, allows SVM to capture relations between similar targets and their compounds and, as such, to 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  $K_t^{\text{kligs}}$  similarity function, then the same function was used as the target-based kernel. For the compound-based kernel function, we used the Tanimoto coefficient (eq 1) as the kernel function for the compounds ( $K_c$ ), as it has been shown to produce good results for building SVM-based SAR models.

Note that a problem with the definitions of the  $K_t^{\text{seq}}$  and  $K_t^{\text{kligs}}$  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.<sup>48</sup> that converts a symmetric matrix into positive semidefinite 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  $K_i^{\text{seq}}$  and  $K_i^{\text{klig}}$  functions.

**2.7. Multi-Assay-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,<sup>23–25</sup> 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  $R_i = \{t_{i_1}, t_{i_2}, \dots, t_{i_m}\}$ . 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  $R$ , 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 learned 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  $c_i$  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

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

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

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

where  $\alpha$  ( $0 \leq \alpha \leq 1$ ) 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 eqs 10 and 11 are based on the  $\text{CWS}_{\alpha}^{\text{sim}}$  and  $\text{CWS}_{\alpha}^{\text{exp}}$  schemes (Section

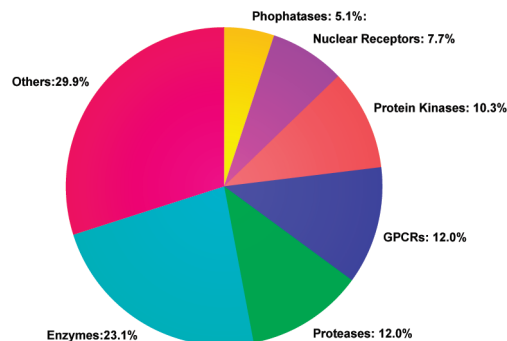


Figure 1. Distribution of protein targets.

2.5.2), respectively. We refer to these three prediction combination schemes as  $\text{MWS}_{\alpha}^{\text{eq}}$ ,  $\text{MWS}_{\alpha}^{\text{sim}}$ , and  $\text{MWS}_{\alpha}^{\text{exp}}$ , respectively.

### 3. MATERIALS

**3.1. Data Sets.** We evaluated the performance of the multi-assay-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 the National Institute of Health's Molecular Libraries Probe Production Centers Network (MLPCN) and are available in PubChem.<sup>49</sup> 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 5-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).<sup>50</sup> 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 of similarity searching. The GF descriptors were generated using the AFGEN<sup>51</sup> program, and they 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<sup>light</sup><sup>52</sup> that implements an efficient soft margin optimization algorithm. In all of our experiments, we used the default parameters



**Table 1.** Performance Improvements of Multi-Assay-Based Semi-Supervised Learning Methods<sup>a</sup>

	$m$	$ U_i $	$ H_i^+ $	$CWS_{\text{none}}$	LS <sub>knn</sub>				$ H_i^+ $	$CWS_{\text{none}}$	LS <sub>LP</sub>			
					CWS <sub>α</sub> <sup>exp</sup>						CWS <sub>α</sub> <sup>exp</sup>			
					0.2	0.5	0.8	0.9			0.2	0.5	0.8	0.9
$K_i^{\text{seq}}$	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%
$K_i^{\text{aligns}}$	1	41	26	<u>1.1%</u>	-0.4%	<u>0.9%</u>	<u>0.8%</u>	<u>0.9%</u>	28	<b>1.8%</b>	<u>0.7%</u>	<u>1.7%</u>	<u>1.3%</u>	<u>0.8%</u>
	3	122	70	<u>0.7%</u>	-1.0%	<u>0.5%</u>	<b>1.2%</b>	<u>0.9%</u>	78	<u>1.3%</u>	<u>0.2%</u>	<u>1.5%</u>	<u>1.5%</u>	<b>1.8%</b>
	5	216	106	-0.5%	-1.5%	-0.2%	<u>0.6%</u>	<u>0.8%</u>	122	<u>0.8%</u>	<u>0.4%</u>	<u>0.8%</u>	<u>0.9%</u>	<u>1.5%</u>
	7	317	134	-0.7%	-1.2%	-0.5%	<u>0.4%</u>	<u>0.5%</u>	243	-0.5%	-0.5%	-0.4%	<u>0.6%</u>	<u>0.9%</u>
	9	432	157	-1.1%	-1.2%	-0.4%	-0.5%	-0.1%	324	-1.3%	-1.0%	-0.9%	0.0%	<u>0.8%</u>
$K_i^{\text{kligs}}$	1	114	61	<u>0.7%</u>	-0.4%	<u>0.8%</u>	<u>0.5%</u>	<u>0.8%</u>	89	-0.3%	-0.6%	0.0%	-0.9%	-0.8%
	3	364	135	-0.4%	-0.9%	-0.5%	-0.2%	<u>0.2%</u>	302	-0.5%	-0.5%	-0.5%	-0.5%	-0.4%
	5	625	179	-1.0%	-1.0%	-0.8%	-0.5%	<u>0.6%</u>	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%

<sup>a</sup> 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<sub>knn</sub> and LS<sub>LP</sub>). The columns labeled 0.2, 0.4, 0.8, and 0.9 correspond to the value of the  $\alpha$  parameter for CWS<sub>α</sub><sup>exp</sup>. The LS<sub>LP</sub> was applied on the five-nearest-neighbor graph of the labeled and unlabeled compounds. The  $K_t^{\text{kligs}}$  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.

for solving the quadratic programming problem and the default regularization parameter *C* that controls the margin width.

**3.4. Evaluation Methodology and Metrics.** The performance of the different methods was evaluated using a 5-fold cross validation framework. For each target *t<sub>i</sub>*, 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.

Note that in the case of the approaches based on semi-supervised learning, some of the newly labeled positive compounds (i.e.,  $H_i^+$ ) may already be in the test set. In such cases, we removed from  $H_i^+$  all these compounds. This allows us to use exactly the same test sets for all the different schemes. This step tends to remove around 2.5% of the newly labeled compounds from  $H_i^+$ .

The quality of the SAR models was measured using the ROC score,<sup>53</sup> which is the normalized area under the curve that plots the true positives against the false positives for different thresholds for classification. Since a 5-fold cross validation was used, the computed ROC scores correspond to the average of the 5-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, which was computed as the ratio of the ROC score of our schemes over that obtained by the baseline method. 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 multi-assay-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 Supporting Information, <http://www-users.cs.umn.edu/~xning/supplementary/>.

**4.1. Performance of the Methods Based on Semi-Supervised Learning.** Table 1 shows the average improvements achieved by the multi-assay-based semi-supervised learning methods over the baseline methods over the entire set of targets in our data set. These results show that for certain parameter combinations the multi-assay-based semi-supervised learning approaches can achieve consistent improvements over the baseline model. The best performance achieved by the multi-assay-based semi-supervised learning approach is an average improvement of 1.8% ( $K_t^{\text{aligns}}$  with LS<sub>LP</sub>).

Comparing the performance achieved by the three target-to-target similarity measures, we see that  $K_t^{\text{aligns}}$  achieves consistently better results, whereas the results achieved by  $K_t^{\text{seq}}$  are consistently the worst. The performance of  $K_t^{\text{kligs}}$  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  $K_t^{\text{aligns}}$  and  $K_t^{\text{kligs}}$  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  $K_t^{\text{kligs}}$  results in a larger number of unlabeled instances (second column of Table 1) than the corresponding set for  $K_t^{\text{aligs}}$ . As a result, the number of positive-labeled instances is larger for  $K_t^{\text{kligs}}$  than for  $K_t^{\text{aligs}}$  (columns labeled as  $|H_i|$ ), which creates more diverse training sets that do not lead to good models. This difference between  $K_t^{\text{aligs}}$  and  $K_t^{\text{kligs}}$  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.

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

Comparing the performance of the different compound weighting schemes ( $\text{CWS}_{\text{none}}$  and  $\text{CWS}_{\alpha}^{\text{exp}}$ ), we see that as the number of unlabeled compounds that are labeled as positive increases,  $\text{CWS}_{\alpha}^{\text{exp}}$  does better than  $\text{CWS}_{\text{none}}$ . 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 of 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  $K_t^{\text{seq}}$  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 multi-assay-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  $K_t^{\text{aligs}}$  and  $K_t^{\text{kligs}}$  kernel functions achieve substantial improvements that range between 2.9 and 7.2%. Moreover, even the  $K_t^{\text{seq}}$  kernel function, which in the context of semi-supervised learning (Table 1) always resulted in lower ROC scores 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,  $K_t^{\text{kligs}}$  achieves the best

**Table 2.** Performance Improvements of the Multi-Assay-Based Multi-Task Learning Methods<sup>a</sup>

	$m$	$\beta$			
		0.1	0.2	0.5	0.8
$K_t^{\text{seq}}$	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%
$K_t^{\text{aligs}}$	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%
$K_t^{\text{kligs}}$	1	4.3%	3.7%	3.1%	2.9%
	3	7.0%	7.1%	4.9%	4.0%
	5	7.0%	<b>7.2%</b>	5.5%	4.1%
	7	6.4%	6.8%	5.3%	3.5%
	9	6.6%	6.9%	5.2%	3.4%

<sup>a</sup> 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  $\beta$  parameter (i.e., weight on the target-based kernel). The  $K_t^{\text{kligs}}$  target-to-target similarity used  $k = 5$ . Bold-faced numbers indicate the best performance of multi-assay-based multi-task learning. Underlined numbers represent schemes with positive improvements.

performance in 14 cases and  $K_t^{\text{aligs}}$  achieves it in 6, whereas  $K_t^{\text{seq}}$  never outperforms the other methods. The best overall performance is achieved by  $K_t^{\text{kligs}}$ , which is a 7.2% improvement over the baseline model. The relatively poor performance of  $K_t^{\text{seq}}$  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  $K_t^{\text{kligs}}$  over  $K_t^{\text{aligs}}$  is somewhat surprising. This is due to the higher diversity among the targets identified by  $K_t^{\text{kligs}}$  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  $K_t^{\text{aligs}}$  and  $K_t^{\text{kligs}}$ , the performance first improves and then degrades. The best performance is usually achieved when 3–5 related targets are used. However, for  $K_t^{\text{seq}}$ , 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  $\beta$  changes (eq 2), we see that they exhibit distinctly different trends. The performance of  $K_t^{\text{aligs}}$  remains largely unchanged as  $\beta$  ranges from 0.1 to 0.8, whereas the performance of  $K_t^{\text{kligs}}$  tends to markedly decrease for higher values of  $\beta$ . Thus, these results indicate that for  $K_t^{\text{kligs}}$ , 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-



**Table 3.** Performance Improvements of Multi-Assay-Based Multi-Ranking Methods<sup>a</sup>

	<i>m</i>	MWS <sub>α</sub> <sup>eq</sup>				MWS <sub>α</sub> <sup>sim</sup>				MWS <sub>α</sub> <sup>exp</sup>			
		0.2	0.5	0.8	0.9	0.2	0.5	0.8	0.9	0.2	0.5	0.8	0.9
<i>K</i> <sub><i>t</i></sub> <sup>seq</sup>	1	−2.7%	<u>1.7%</u>	<u>1.8%</u>	<u>1.3%</u>	−2.7%	<u>1.7%</u>	<u>1.8%</u>	<u>1.3%</u>	−3.0%	<u>1.7%</u>	<u>2.0%</u>	<u>1.3%</u>
	3	−13.5%	<u>0.7%</u>	<u>2.3%</u>	<u>1.5%</u>	−11.8%	<u>1.6%</u>	<u>2.5%</u>	<u>1.4%</u>	−12.0%	<u>0.2%</u>	<u>3.1%</u>	<u>1.9%</u>
	5	−19.8%	−5.0%	<u>2.1%</u>	<u>1.5%</u>	−18.6%	−2.8%	<u>2.1%</u>	<u>2.0%</u>	−17.0%	−1.5%	<u>3.0%</u>	<u>2.1%</u>
	7	−20.2%	−9.8%	<u>1.7%</u>	<u>1.2%</u>	−19.4%	−6.3%	<u>2.1%</u>	<u>1.7%</u>	−18.0%	−2.1%	<u>2.9%</u>	<u>2.1%</u>
	9	−23.9%	−17.8%	<u>2.1%</u>	<u>1.3%</u>	−23.8%	−14.6%	<u>2.7%</u>	<u>1.6%</u>	−22.1%	−2.8%	<u>2.8%</u>	<u>2.1%</u>
<i>K</i> <sub><i>t</i></sub> <sup>aligs</sup>	1	<u>1.3%</u>	<u>2.9%</u>	<u>2.4%</u>	<u>1.7%</u>	<u>1.3%</u>	<u>2.9%</u>	<u>2.4%</u>	<u>1.7%</u>	<u>1.0%</u>	<u>2.9%</u>	<u>2.2%</u>	<u>1.7%</u>
	3	−6.6%	<u>3.8%</u>	<u>2.5%</u>	<u>1.8%</u>	−5.7%	<u>4.2%</u>	<u>2.5%</u>	<u>1.9%</u>	−4.9%	<u>4.2%</u>	<u>3.1%</u>	<u>2.6%</u>
	5	−12.1%	−0.4%	<u>2.0%</u>	<u>1.8%</u>	−11.7%	<u>0.5%</u>	<u>2.1%</u>	<u>1.9%</u>	−10.2%	<u>2.7%</u>	<u>4.0%</u>	<u>2.9%</u>
	7	−12.9%	−4.6%	<u>3.0%</u>	<u>1.9%</u>	−12.7%	−3.4%	<u>3.1%</u>	<u>1.8%</u>	−11.0%	<u>2.1%</u>	<u>4.0%</u>	<u>3.0%</u>
	9	−13.1%	−7.8%	<u>4.6%</u>	<u>2.0%</u>	−13.2%	−7.1%	<u>4.4%</u>	<u>2.0%</u>	−11.3%	<u>2.4%</u>	<u>3.9%</u>	<u>3.0%</u>
<i>K</i> <sub><i>t</i></sub> <sup>kligs</sup>	1	<u>0.8%</u>	<u>4.0%</u>	<u>4.2%</u>	<u>3.1%</u>	<u>0.8%</u>	<u>4.0%</u>	<u>4.2%</u>	<u>3.1%</u>	<u>1.0%</u>	<u>4.0%</u>	<u>4.0%</u>	<u>3.1%</u>
	3	−5.1%	<u>5.1%</u>	<u>5.3%</u>	<u>3.2%</u>	−4.5%	<u>5.8%</u>	<u>5.5%</u>	<u>3.2%</u>	−3.3%	<u>6.4%</u>	<u>6.0%</u>	<u>4.8%</u>
	5	−10.2%	<u>1.4%</u>	<u>6.3%</u>	<u>4.0%</u>	−9.7%	<u>2.4%</u>	<b>6.4%</b>	<u>4.1%</u>	−6.2%	<u>6.6%</u>	<b>7.0%</b>	<u>5.4%</u>
	7	−13.6%	−5.4%	<b>6.5%</b>	<u>4.4%</u>	−13.4%	−4.6%	<u>6.1%</u>	<u>4.4%</u>	−11.0%	<u>6.0%</u>	<b>7.0%</b>	<u>5.5%</u>
	9	−16.2%	−10.6%	<u>5.6%</u>	<u>3.4%</u>	−16.1%	−9.8%	<u>5.9%</u>	<u>3.6%</u>	−14.2%	<u>5.2%</u>	<b>7.0%</b>	<u>5.5%</u>

<sup>a</sup> 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  $\alpha$  parameter for MWS<sub>α</sub><sup>eq</sup>, MWS<sub>α</sub><sup>sim</sup>, and MWS<sub>α</sub><sup>exp</sup>, respectively. The *K*<sub>*t*</sub><sup>kligs</sup> 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.

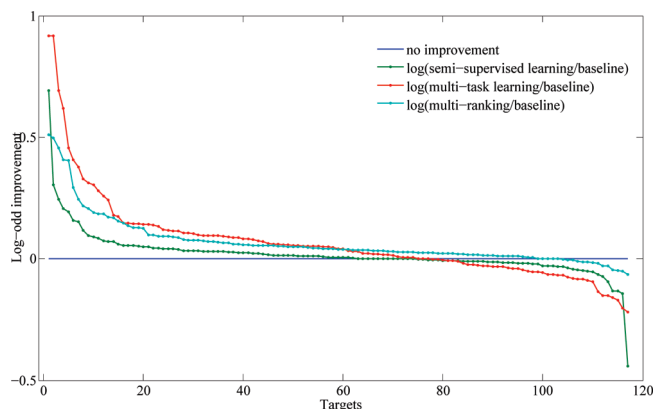
**Table 4.** Summary of the Performance Improvements of the Different Multi-Assay-Based Methods<sup>a</sup>

methods	target-to-target similarity	compound labeling	weighting scheme	Weight	<i>m</i>	best average improvement	ROC	% improved targets	<i>p</i> -value
semi-supervised learning	<i>K</i> <sub><i>t</i></sub> <sup>aligs</sup>	LS <sub>LP</sub>	CWS <sub>α</sub> <sup>exp</sup>	$\alpha = 0.9$	3	<u>1.8%</u>	0.66	54%	$5.69 \times 10^{-2}$
multi-task learning	<i>K</i> <sub><i>t</i></sub> <sup>kligs</sup>	—	—	$\beta = 0.2$	5	<b>7.2%</b>	0.70	63%	$6.39 \times 10^{-5}$
multi-ranking	<i>K</i> <sub><i>t</i></sub> <sup>kligs</sup>	—	MWS <sub>α</sub> <sup>exp</sup>	$\alpha = 0.8$	5	<u>7.0%</u>	0.70	85%	$7.06 \times 10^{-11}$

<sup>a</sup> In this table, *m* is the number of related targets, best average improvement is the geometric mean achieved for each of the multi-assay-based methods under the parameter combination shown in the corresponding row, ROC is the average area under the ROC curved achieved by each scheme, % improved targets is the percentage of the 117 targets for which the multi-assay-based method resulted in better performance. The *p*-value of the statistical significance test, using the paired students *t* test of the results achieved by each scheme over the baseline approach, is shown in the column labeled *p*-value. The average ROC for baseline model was 0.65. Bold-faced numbers indicate the best performance over all multi-assay-based methods. Underlined numbers represent the schemes with positive improvements.

to-target similarity measures are consistent with the results obtained using the multi-assay-based multi-task learning method. *K*<sub>*t*</sub><sup>kligs</sup> tends to perform the best, with some schemes achieving an average improvement of 7.0%, whereas *K*<sub>*t*</sub><sup>seq</sup> does relatively worse 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 the target-to-target similarity or kernel functions. Comparing the three prediction combination schemes MWS<sub>α</sub><sup>eq</sup>, MWS<sub>α</sub><sup>sim</sup>, and MWS<sub>α</sub><sup>exp</sup>, we see that, on average, MWS<sub>α</sub><sup>exp</sup> performs the best and is followed by MWS<sub>α</sub><sup>sim</sup>, and MWS<sub>α</sub><sup>eq</sup> 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 MWS<sub>α</sub><sup>exp</sup> 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

**Figure 2.** Improvement log ratio distribution.

large *m* is because less similar models make less reliable predictions, and thus, combining them will not introduce any benefits.

## 5. DISCUSSION

**5.1. Overall Performance.** Table 4 summarizes the best results achieved by the three multi-assay-based schemes developed in this work, whereas Figure 2 presents a finer grain view of these results by plotting the log ratio of the ROC score achieved by each of them over the ROC score

**Table 5.** ROC Scores for Baseline, Semi-Supervised Learning, Multi-Task Learning, and Multi-Ranking

target ID	baseline	semi-supervised learning	multi-task learning	multi-ranking	target ID	baseline	semi-supervised learning	multi-task learning	multi-ranking
1	0.300	0.600	0.750	0.500	60	0.664	0.674	0.708	0.705
2	0.311	0.200	0.578	0.467	61	0.666	0.684	0.739	0.718
3	0.311	0.422	0.778	0.511	62	0.669	0.641	0.651	0.686
4	0.392	0.404	0.531	0.445	63	0.669	0.661	0.649	0.669
5	0.400	0.400	0.800	0.600	64	0.672	0.671	0.548	0.659
6	0.400	0.511	0.556	0.511	65	0.675	0.669	0.711	0.691
7	0.412	0.500	0.650	0.650	66	0.675	0.704	0.708	0.740
8	0.475	0.500	0.713	0.637	67	0.676	0.677	0.713	0.701
9	0.511	0.528	0.507	0.519	68	0.676	0.726	0.622	0.730
10	0.525	0.575	0.694	0.653	69	0.685	0.677	0.689	0.704
11	0.533	0.600	0.778	0.644	70	0.686	0.671	0.684	0.693
12	0.536	0.558	0.582	0.573	71	0.686	0.739	0.624	0.706
13	0.541	0.572	0.465	0.570	72	0.687	0.696	0.716	0.705
14	0.544	0.584	0.704	0.600	73	0.688	0.640	0.766	0.744
15	0.547	0.539	0.530	0.541	74	0.695	0.689	0.641	0.700
16	0.551	0.600	0.608	0.653	75	0.696	0.664	0.680	0.736
17	0.555	0.562	0.587	0.630	76	0.699	0.698	0.663	0.700
18	0.561	0.579	0.667	0.602	77	0.700	0.700	0.800	0.750
19	0.561	0.589	0.450	0.613	78	0.702	0.682	0.767	0.757
20	0.567	0.582	0.572	0.579	79	0.702	0.725	0.740	0.742
21	0.576	0.565	0.633	0.609	80	0.707	0.731	0.759	0.742
22	0.576	0.592	0.664	0.672	81	0.710	0.688	0.748	0.742
23	0.586	0.583	0.563	0.592	82	0.711	0.711	0.822	0.844
24	0.596	0.604	0.684	0.654	83	0.712	0.744	0.784	0.806
25	0.600	0.562	0.675	0.662	84	0.720	0.792	0.736	0.776
26	0.600	0.578	0.511	0.644	85	0.728	0.768	0.872	0.756
27	0.601	0.601	0.636	0.619	86	0.730	0.732	0.723	0.727
28	0.606	0.600	0.567	0.604	87	0.731	0.719	0.756	0.746
29	0.609	0.609	0.687	0.638	88	0.732	0.719	0.673	0.721
30	0.610	0.641	0.707	0.698	89	0.733	0.711	0.822	0.800
31	0.611	0.604	0.641	0.631	90	0.736	0.759	0.819	0.777
32	0.613	0.713	0.837	0.737	91	0.737	0.730	0.644	0.704
33	0.616	0.584	0.691	0.653	92	0.738	0.863	0.850	0.888
34	0.619	0.623	0.592	0.601	93	0.743	0.762	0.722	0.749
35	0.621	0.617	0.607	0.635	94	0.747	0.788	0.762	0.772
36	0.624	0.596	0.627	0.661	95	0.752	0.736	0.816	0.760
37	0.626	0.607	0.528	0.649	96	0.755	0.784	0.649	0.718
38	0.626	0.626	0.625	0.626	97	0.756	0.747	0.715	0.766
39	0.631	0.637	0.651	0.640	98	0.756	0.772	0.827	0.799
40	0.631	0.650	0.635	0.656	99	0.756	0.774	0.832	0.788
41	0.635	0.643	0.610	0.663	100	0.758	0.791	0.827	0.788
42	0.636	0.644	0.700	0.662	101	0.761	0.761	0.771	0.761
43	0.637	0.641	0.672	0.660	102	0.762	0.762	0.737	0.788
44	0.637	0.650	0.597	0.681	103	0.768	0.772	0.744	0.732
45	0.637	0.663	0.812	0.737	104	0.769	0.773	0.764	0.786
46	0.638	0.638	0.700	0.700	105	0.770	0.766	0.784	0.786
47	0.641	0.555	0.608	0.633	106	0.772	0.759	0.788	0.812
48	0.641	0.630	0.680	0.659	107	0.772	0.764	0.758	0.785
49	0.643	0.635	0.682	0.667	108	0.778	0.756	0.711	0.756
50	0.645	0.641	0.677	0.673	109	0.785	0.793	0.782	0.793
51	0.645	0.648	0.656	0.655	110	0.795	0.806	0.829	0.817
52	0.647	0.647	0.666	0.639	111	0.797	0.791	0.859	0.837
53	0.650	0.800	0.750	0.800	112	0.797	0.809	0.843	0.816
54	0.651	0.666	0.651	0.671	113	0.799	0.807	0.770	0.811
55	0.654	0.595	0.611	0.650	114	0.800	0.700	0.750	0.750
56	0.659	0.671	0.674	0.675	115	0.800	0.700	0.800	0.850
57	0.660	0.681	0.714	0.693	116	0.800	0.760	0.760	0.808
58	0.660	0.682	0.612	0.664	117	0.800	0.800	0.800	0.800
59	0.662	0.703	0.767	0.697					

achieved by the baseline model for the 117 targets in our data set. Note that for each multi-assay-based scheme, the results in Figure 2 are presented in a nonincreasing order according to these log ratios. The actual ROC scores for the 117 targets and the four schemes (baseline and multi-assay-based methods) are shown in Table 5.

These results provide strong evidence that the multi-assay-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, in terms of average ROC, multi-task learning and multi-ranking perform comparably and achieve the best overall results, whereas in terms of performance consistency over the baseline approach, the multi-ranking method performs the best leading to better ROC scores for 99 out of the 117 targets (85% of the targets). Moreover,

**Table 6.** Performance Improvements of the Different Multi-Assay-Based Methods in the Presence of False Positives<sup>a</sup>

method	actual inactives as false positives				nonactives as false positives			
	1%	2%	5%	10%	1%	2%	5%	10%
semi-supervised learning	1.8% ( $5.88 \times 10^{-2}$ )	1.6% ( $1.32 \times 10^{-1}$ )	1.1% ( $4.29 \times 10^{-1}$ )	0.7% ( $4.58 \times 10^{-1}$ )	1.8% ( $5.47 \times 10^{-2}$ )	1.6% ( $1.35 \times 10^{-1}$ )	1.0% ( $3.97 \times 10^{-1}$ )	0.7% ( $5.49 \times 10^{-1}$ )
multi-task learning	7.0% ( $2.23 \times 10^{-4}$ )	6.4% ( $1.23 \times 10^{-3}$ )	3.5% ( $2.05 \times 10^{-2}$ )	3.4% ( $2.60 \times 10^{-2}$ )	6.9% ( $3.61 \times 10^{-4}$ )	6.2% ( $2.46 \times 10^{-3}$ )	3.3% ( $3.75 \times 10^{-2}$ )	3.2% ( $1.96 \times 10^{-2}$ )
multi-ranking	6.9% ( $1.14 \times 10^{-10}$ )	5.8% ( $2.82 \times 10^{-8}$ )	2.0% ( $1.12 \times 10^{-4}$ )	1.5% ( $5.79 \times 10^{-2}$ )	6.9% ( $1.38 \times 10^{-10}$ )	5.5% ( $3.15 \times 10^{-8}$ )	1.9% ( $1.21 \times 10^{-4}$ )	1.5% ( $5.84 \times 10^{-2}$ )

<sup>a</sup> For each learning method in the rows, the numbers on top (percentage numbers) are the performance improvements, and the numbers on bottom (powered numbers) are the *p*-values from paired *t*-test.

the performance gains achieved by multi-task learning and multi-ranking are statistically significant using a paired *t*-test with *p*-values of  $1.3 \times 10^{-5}$  and  $8.6 \times 10^{-14}$ , respectively. 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, which accounts for the characteristics of the ligand binding sites, whereas the former does not. Third, the three multi-assay-based methods behave differently for the two ligand-based target-to-target similarity functions. Semi-supervised learning performs best for  $K_t^{\text{aligs}}$ , whereas the other two perform better for  $K_t^{\text{kligs}}$ . As discussed in,  $K_t^{\text{kligs}}$  tends to select targets that have a large number of ligands. In the context of semi-supervised learning methods, this leads to a large number of unlabeled instances, which is the reason behind the lower performance of  $K_t^{\text{kligs}}$  over  $K_t^{\text{aligs}}$ . However, in the case of the methods based on multi-task learning and multi-ranking, this property of  $K_t^{\text{kligs}}$  actually leads to improved performance. This is because the targets selected by  $K_t^{\text{kligs}}$  tend to contain more diverse sets of compounds than those selected by  $K_t^{\text{aligs}}$  (the average pairwise compound similarity of  $K_t^{\text{aligs}}$ 's five most similar targets was 0.0138, whereas the corresponding similarity for  $K_t^{\text{kligs}}$  was only 0.0071) and consequently, there is a higher degree of diversity among the set of targets that are being selected by  $K_t^{\text{kligs}}$ . 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.<sup>23–25</sup>

**5.2. Performance on Actual Inactives.** One of the decisions that was made in building the various SAR models for each target  $t_i$  was to ignore the inactive compounds that may be available for  $t_i$  (see the discussion in Section 2.2 for the reasons underlying this decision) and to use instead a random subset of the ligands of the other targets as the set of inactive compounds for  $t_i$  during model learning (following a learning from only positive and unlabeled instances framework).<sup>31</sup> To assess the extent to which this framework for building the multi-assay-based models can still lead to improvements over the baseline models when the actual inactives are being predicted, we performed an experiment in which the test set consisted of both actual active and inactive compounds. These experiments showed that the multi-assay-based models can still improve over the baseline models, achieving an average ROC improvement of 2.2, 3.9,

and 2.4% for the semi-supervised learning, multi-task learning, and multi-ranking, respectively.

**5.3. False Positive Sensitivity.** To assess the robustness of the models in the presence of false positives, we performed a series of experiments in which different multi-assay-based and baseline models were built in which a certain fraction of false positives were included as additional positive training instances. We used two different approaches for selecting the false positives. The first selected the false positives from the actual inactives of each target, whereas the second selected the false positives from the ligands of the other targets that are not true positives for the target under consideration (i.e., the same approach that was used to generate the negative instances for the results presented in Section 4). The resulting models were then tested on compounds consisting of confirmed and randomly selected actives from other targets as inactives. To ensure that the comparisons with the baseline method was done fairly, the same set of training sets (i.e., with false positives) were also used to build the baseline models.

The average ROC improvements of the best parameter combination of the multi-assay-based models over the baseline model for these sets of experiments are shown in Table 6 (this table is the analog of Table 4). Results for the different number of false positives as a percentage of the actual number of positives are presented ranging from 1 up to 10%. These results show that even in the presence of false positives, the multi-assay-based approaches lead to improvements over the baseline approach. For small percentages of false positives (1–2%), the relative gains achieved by the multi-task learning and multi-ranking approaches remain considerable. However, as the percentage of false positives increases, the relative performance gains over the baseline approach decreases. For example, at 10% of false positives, the improvements achieved by multi-task learning and multi-ranking drops to around 3.4 and 1.5%, respectively. This suggests that, in the presence of a considerable number of false positive instances, these approaches fail to identify a good set of related targets and/or utilize their activity information to improve the quality of the SAR models. Finally, these results suggest that among the multi-assay-based methods, multi-task learning is more robust in dealing with the presence of a considerable number of false positives.

**5.4. Comparison with Chemogenomics-Based Approaches.** 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

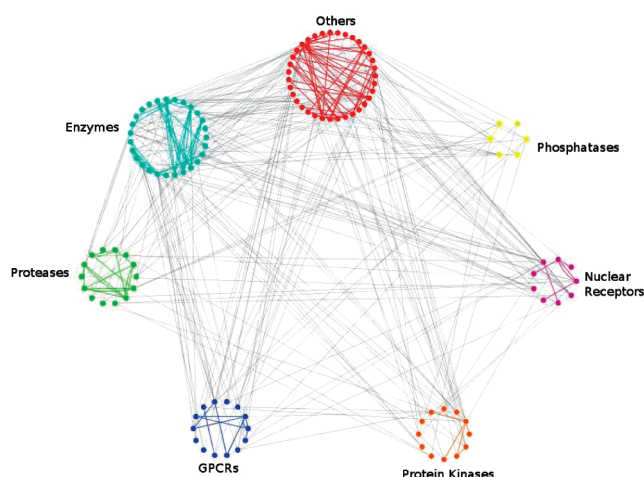


**Table 7.** Performance of Chemogenomics- And Multi-Assay-Based Approaches Relative to the Baseline Models<sup>a</sup>

family	scheme	multi-task learning				multi-ranking				
		target-to-target similarity	$\beta$	$m$	imprvmnt	target-to-target similarity	weighting scheme	$\alpha$	$m$	imprvmnt
phosphatases	ChmGnmics	$K_t^{\text{aligs}}$	0.2	6	−5.2%	$K_t^{\text{eq}}$	$\text{MWS}_{\alpha}^{\text{sim}}$	0.9	6	0.2%
	MABfamily	$K_t^{\text{kligs}}$	0.2	1	1.2%	$K_t^{\text{kligs}}$	$\text{MWS}_{\alpha}^{\text{exp}}$	0.9	3	−2.3%
	MABglobal	$K_t^{\text{aligs}}$	0.2	1	<b>6.9%</b>	$K_t^{\text{aligs}}$	$\text{MWS}_{\alpha}^{\text{exp}}$	0.8	3	6.5%
nuclear receptors	ChmGnmics	$K_t^{\text{kligs}}$	0.2	9	14.6%	$K_t^{\text{kligs}}$	$\text{MWS}_{\alpha}^{\text{exp}}$	0.8	9	11.8%
	MABfamily	$K_t^{\text{kligs}}$	0.2	3	14.0%	$K_t^{\text{kligs}}$	$\text{MWS}_{\alpha}^{\text{exp}}$	0.5	3	13.3%
	MABglobal	$K_t^{\text{kligs}}$	0.2	5	<b>18.1%</b>	$K_t^{\text{kligs}}$	$\text{MWS}_{\alpha}^{\text{exp}}$	0.8	5	10.6%
protein kinases	ChmGnmics	$K_t^{\text{aligs}}$	0.5	12	4.3%	$K_t^{\text{aligs}}$	$\text{MWS}_{\alpha}^{\text{exp}}$	0.8	12	8.2%
	MABfamily	$K_t^{\text{kligs}}$	0.2	5	11.3%	$K_t^{\text{kligs}}$	$\text{MWS}_{\alpha}^{\text{sim}}$	0.8	7	9.8%
	MABglobal	$K_t^{\text{kligs}}$	0.2	1	<b>15.3%</b>	$K_t^{\text{kligs}}$	$\text{MWS}_{\alpha}^{\text{exp}}$	0.8	7	14.5%
GPCRs	ChmGnmics	$K_t^{\text{aligs}}$	0.2	14	−3.6%	$K_t^{\text{kligs}}$	$\text{MWS}_{\alpha}^{\text{exp}}$	0.8	14	1.4%
	MABfamily	$K_t^{\text{aligs}}$	0.2	1	1.4%	$K_t^{\text{aligs}}$	$\text{MWS}_{\alpha}^{\text{sim}}$	0.8	7	2.6%
	MABglobal	$K_t^{\text{aligs}}$	0.2	3	<b>6.8%</b>	$K_t^{\text{kligs}}$	$\text{MWS}_{\alpha}^{\text{exp}}$	0.8	1	3.2%
proteases	ChmGnmics	$K_t^{\text{aligs}}$	0.2	14	0.8%	$K_t^{\text{kligs}}$	$\text{MWS}_{\alpha}^{\text{exp}}$	0.8	14	5.1%
	MABfamily	$K_t^{\text{kligs}}$	0.2	1	6.7%	$K_t^{\text{kligs}}$	$\text{MWS}_{\alpha}^{\text{exp}}$	0.8	3	6.4%
	MABglobal	$K_t^{\text{kligs}}$	0.2	5	<b>12.1%</b>	$K_t^{\text{kligs}}$	$\text{MWS}_{\alpha}^{\text{exp}}$	0.8	7	9.3%
enzymes	ChmGnmics	$K_t^{\text{aligs}}$	0.2	27	−6.6%	$K_t^{\text{kligs}}$	$\text{MWS}_{\alpha}^{\text{exp}}$	0.8	27	1.3%
	MABfamily	$K_t^{\text{aligs}}$	0.2	1	0.4%	$K_t^{\text{kligs}}$	$\text{MWS}_{\alpha}^{\text{exp}}$	0.8	7	1.3%
	MABglobal	$K_t^{\text{kligs}}$	0.2	5	2.7%	$K_t^{\text{kligs}}$	$\text{MWS}_{\alpha}^{\text{exp}}$	0.8	9	<b>2.9%</b>

<sup>a</sup> In this table,  $\beta$  is the weight on target similarity for multi-assay-based multi-task learning method,  $\alpha$  is the weight for  $\text{MWS}_{\alpha}^{\text{eq}}$ ,  $\text{MWS}_{\alpha}^{\text{sim}}$ , and  $\text{MWS}_{\alpha}^{\text{exp}}$ ,  $m$  is either the number of related targets (multi-assay-based approaches) or the size of the protein family (chemogenomics-based approaches), imprvmnt is the performance of certain multi-assay-based scheme under corresponding combination of parameters for each protein family. ChemGnmics denotes the results obtained by the chemogenomics-based approach, MABfamily denotes the results obtained by the family-focused multi-assay-based approach, and MABglobal denotes the results obtained by the actual multi-assay-based approach. Bold-faced numbers indicate the best performing scheme for each protein family.

context of the methods introduced in this paper, these chemogenomics-based approaches can be viewed as a special case of the multi-assay-based models in which all the proteins of the same family become the set of related proteins. Table 7 shows the performance gains achieved by the chemogenomics- and multi-assay-based approaches over the baseline models on the six protein families in our data set that contain at least four members (i.e., the proteins in the “other” class of 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 multi-assay-based approach (labeled “MABfamily” and “MABglobal”). The MABglobal method corresponds to the multi-assay-based schemes that were described in Section 2, whereas the MABfamily 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 multi-assay-based approaches. Averaged over the 82 proteins in these six families, the multi-assay-based approaches achieve a 4.33% improvement over the chemogenomics-based approaches (best multi-assay-based scheme vs best chemogenomics-based scheme). In addition, comparing the performance achieved by the MABfamily variant of the multi-assay-based methods, we see that they perform 0.9% better than the chemogenomics-based approaches and 3.3% worse than the actual multi-assay-based approaches (MABglobal). These results show that higher performance gains can be obtained by not utilizing the activity information from all the proteins in the family

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

(MABfamily vs ChmGnmics) and that even further gains can be achieved by utilizing activity information from proteins of different families (MABglobal vs MABfamily).

To illustrate the cross-family nature of the multi-assay-based methods, Figure 3 shows the set of related proteins for the different proteins within and across the different families ( $K_t^{\text{kligs}}$  and  $m = 3$ ). This figure shows that, for nearly all protein targets, a fair number of their related targets (66.5%) comes from targets that belong to other families and includes 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, on multi-task learning that attempt to improve the quality of the SAR model by transferring information learned from different targets, and 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 data set of 117 protein targets has 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 both 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, these methods, by selecting the targets from which activity information will be utilized from the entire data set, outperform approaches based on chemogenomics that utilize activity information of protein targets belonging to the same family as that under consideration.

#### ACKNOWLEDGMENT

This work was supported by the National Science Foundation IIS-0431135, the National Institutes of Health RLM008713A, and the Digital Technology Center at the University of Minnesota.

**Supporting Information Available:** Results for multi-assay-based semi-supervised, multi-assay-based multi-task learning, methods, multi-assay-based multi-ranking methods. Information can be found at <http://www-users.cs.umn.edu/~xning/supplementary/>. This material is available free of charge via the Internet at <http://pubs.acs.org>.

#### REFERENCES AND NOTES

- Hansch, C.; Maolney, P. P.; Fujita, T.; Muir, R. M. *Nature* **1962**, *194*, 178–180.
- Hansch, C.; Muir, R. M.; Fujita, T.; Maloney, C. F.; Streich, M. J. *Am. Chem. Soc.* **1963**, *85*, 2817–1824.
- Bravi, G.; Green, E. G. D.; Hann, V.; Mike, M. Modelling Structure-Activity Relationship. In *Virtual Screening for Bioactive Molecules*, Volume 10; Bohm, H., Schneider, G., Eds.; Wiley-VCH: New York, 2000.
- Agrafiotis, D.; Bandyopadhyay, D.; Wegner, J.; van Vlijmen, H. J. *Chem. Inf. Model.* **2007**, *47*, 1279–1293.
- Warmuth, M. K.; Liao, J.; Ratsch, G.; Mathieson, M.; Putta, S.; Lemmen, C. J. *Chem. Inf. Comput. Sci.* **2003**, *43*, 667–673.
- Frye, S. *Chemistry and Biology* **1999**, R3–R7.
- 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.
- Klabunde, T. Br. J. Pharmacol. **2007**, *152*, 5–7.
- Bock, J.; Gough, D. J. *Chem. Inf. Model.* **2005**, *45*, 1402–1414.
- Lapins, M.; Prusis, P.; Uhlen, S.; Wikberg, J. E. S. *Bioinformatics* **2005**, *21*, 4289–4296.
- Lindström, A.; Pettersson, F.; Almqvist, F.; Berglund, A.; Kihlberg, J.; Linusson, A. J. *Chem. Inf. Model.* **2006**, *46*, 1154–1167.
- Jacob, L.; Hoffmann, B.; Stoven, V.; Vert, J.-P. *BMC Bioinformatics* **2008**, *9*, 363.
- Strömbergsson, H.; Daniluk, P.; Kryshchovych, A.; Fidelis, K.; Wikberg, J. E. S.; Kleywegt, G. J.; Hvidsten, T. R. J. *Chem. Inf. Model.* **2008**, *48*, 2278–2288.
- Deng, Z.; Chuaqui, C.; Singh, J. J. *Med. Chem.* **2004**, *47*, 337–344.
- Weill, N.; Rognan, D. J. *Chem. Inf. Model.* **2009**, *49*, 1049–1062.
- Geppert, H.; Humrich, J.; Stumpfe, D.; Gartner, T.; Bajorath, J. J. *Chem. Inf. Model.* **2009**, *49*, 767–779.
- Erhan, D.; L'Heureux, P.-J.; Yue, S. Y.; Bengio, Y. J. *Chem. Inf. Model.* **2006**, *46*, 626–635.
- Semi-Supervised Learning*; Chapelle, O., Schölkopf, B., Zien, A., Eds.; MIT Press: Cambridge, MA, 2006.
- Thurn, S. Is learning the n-th thing any easier than learning the first? In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, 1996.
- Caruana, R. A. Multitask learning: A knowledge-based source of inductive bias. In *Proc. Int. Conf. Mach. Learn.*; Morgan Kaufmann: MA, 1993.
- Evgeniou, T.; Micchelli, C. A.; Pontil, M. J. *Mach. Learn. Res.* **2005**, *6*, 615–637.
- Bonilla, E.; Agakov, F.; Williams, C. Kernel Multi-task Learning using Task-specific Features. In *Proc. Int. Conf. Artificial Intelligence and Statistics*; Omnipress: San Juan, Puerto Rico, 2007.
- Swanson, R.; Tsai, J. J. *Bacteriol.* **2003**, *185*, 3990–3993.
- Kuncheva, L. I.; Whitaker, C. J. *Mach. Learn.* **2003**, *51*, 181–207.
- Shipp, C. A.; Kuncheva, L. I. *Inf. Fusion* **2002**, *3*, 135–148.
- P. Willett, J.; Downs, G. M. J. *Chem. Inf. Comput. Sci.* **1998**, *38*, 983–997.
- Vapnik, V. *Statistical Learning Theory*; John Wiley: New York, 1998.
- Yu, H.; Han, J.; Chang, K. PEBL: Positive example based learning for Web page classification using SVM. In *Proc. Int. Conf. Know. Disc. Data Mining*; ACM: New York, NY, USA, 2002.
- Liu, B.; Dai, Y.; Li, X.; Lee, W.; Yu, P. Building text classifiers using positive and unlabeled examples. In *Proc. Int. Conf. Data Mining*; IEEE Computer Society Press: Florida, USA, 2003.
- Wang, C.; Ding, C.; Meraz, R. F.; Holbrook, S. R. *Bioinformatics* **2006**, *22*, 2590–2596.
- Elkan, C.; Noto, K. Learning classifiers from only positive and unlabeled data. In *Proc. Int. Conf. Know. Disc. Data Mining*; ACM: New York, 2008.
- Davies, E. K.; Briant, C. Combinatorial Chemistry Library Design Using Pharmacophore Diversity. In *Molecular Diversity and Combinatorial Chemistry: Libraries and Drug Discovery*, Volume 118; American Chemical Society: Washington, DC, **1996**, 309–316.
- Smith, T. F.; Waterman, M. S. J. *Mol. Biol.* **1981**, *147*, 195–197.
- Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. *Nucleic Acids Res.* **1997**, *25*, 3389–3402.
- Heger, A.; Holm, L. *Bioinformatics* **2001**, *17*, 272–279.
- Rangwala, H.; Karypis, G. *Bioinformatics* **2005**, *21*, 4239–4247.
- Zhu, X. “Semi-Supervised Learning Literature Survey”, Technical Report, Computer Sciences, University of Wisconsin: Madison, WI, 2005.
- 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.
- Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. J. *Med. Chem.* **2002**, *45*, 4350–4358.
- Weston, J.; Elisseeff, A.; Zhou, D.; Leslie, C.; Noble, W. S. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 6559–6563.
- Tetko, I. V.; Tanchuk, V. Y. J. *Chem. Inf. Comput. Sci.* **2002**, *42*, 1136–1145.
- Tetko, I. V.; Jaroszewicz, I.; Platts, J. A.; Kuduk-Jaworska, J. J. *Inorg. Biochem.* **2008**, *102*, 1424–1437.
- Varnek, A.; Gaudin, C.; Marcou, G.; Baskin, I.; Pandey, A. K.; Tetko, I. V. J. *Chem. Inf. Model.* **2009**, *49*, 133–144.
- Landkriet, G. R.; Deng, M.; Cristianini, N.; Jordan, M. I.; Noble, W. S. Kernel-based data fusion and its application to protein function prediction in yeast. In *Proc. Pac. Symp. Biocomput.* World Scientific Publishing: Berkeley, CA, 2004.
- Sonnenburg, S.; Ratsch, G.; Schafer, C. A General and Efficient Multiple Kernel Learning Algorithm. In *Proc. Neural Inf. Proc. Sys.*; British Columbia, Canada, 2005.
- Tsang, I. W.; Kwok, J. T. *IEEE Trans. Neural Networks* **2006**, *17*, 48–58.
- Rangwala, H.; Karypis, G. fRMSDAlign: Protein Sequence Alignment Using Predicted Local Structure Information for Pairs with Low Sequence Identity. In *Proc. Asia-Pacific Bioinformatics Conf.* Volume 6; Brazma, A., Miyano, S., Akutsu, T., Eds.; Imperial College Press: Kyoto, Japan, 2008.
- Saigo, H.; Vert, J.-P.; Ueda, N.; Akutsu, T. *Bioinformatics* **2004**, *20*, 1682–1689.
- Austin, C. P.; Brady, L. S.; Insel, T. R.; Collins, F. S. *Science* **2004**, *306*, 1138–1139.
- Wale, N.; Watson, I. A.; Karypis, G. *Knowledge and Information Systems* **2008**, *14*, 347–375.
- Karypis, G. AFGEN2.0, Technical Report 08-020, Department of Computer Science & Engineering, University of Minnesota: Minneapolis, MN; 2007 [www.cs.umn.edu/~karypis](http://www.cs.umn.edu/~karypis).
- Joachims, T. *Making Large-Scale SVM Learning Practical* **1999**, 169–184.
- Fawcett, T. *ROC Graphs: Notes and Practical Considerations for Researchers*, HP Lab Technical Report HPL-2003-4; Hewlett Packard: Palo Alto, CA, 2003.