JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

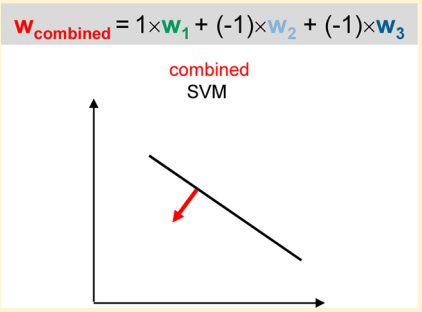
pubs.acs.org/jcim

Prediction of Compounds with Closely Related Activity Profiles Using Weighted Support Vector Machine Linear Combinations

Kathrin Heikamp and Jürgen Bajorath*

Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Dahlmannstr. 2, D-53113 Bonn, Germany

ABSTRACT: Using support vector machine (SVM) ranking, a complex multi-class prediction task has been investigated involving sets of compounds that were active against related targets and represented all possible combinations of single-, dual-, and triple-target activities. Standard SVM models were not capable of differentiating compounds with overlapping yet distinct activity profiles. To address this problem, we designed differentially weighted SVM linear combinations that were found to preferentially detect compounds with desired activity profiles and deprioritize others. Hence, combining independently derived SVM models using negative and positive linear weighting factors balanced relative contributions from individual reference sets and successfully distinguished between compounds with overlapping activity profiles.



INTRODUCTION

Predicting biological activities of small molecules from chemical structure is one of the major focal points of the chemoinformatics field. For this purpose, a plethora of computational methodologies have been introduced. In recent years, machine learning approaches have become increasingly popular for activity predictions, especially Bayesian classifiers[1−4] and support vector machines (SVMs).[5−9] These supervised learning methods have typically been used for binary classification and prediction of class labels of compounds (e.g., active vs inactive) focusing on different compound activity classes.[2,4,8,9]

Given the increasing interest in chemogenomics,[10,11] these machine learning approaches have also been considered for complex activity predictions. In chemogenomics, the systematic analysis of compound-target annotations and the study of multi-target activities of small molecules take center stage. Accordingly, machine learning approaches have been used for applications such as computational profiling of compounds against arrays of classifiers for individual targets,[12] prediction of ligand−receptor pairings,[13,14] or searching for target-selective compounds.[15] For selectivity predictions, SVM modeling was carried out to distinguish target-selective compounds from others that were active against multiple members of a given target family and also from inactive compounds.

In machine learning terms, the chemogenomics-oriented applications outlined above translate into multi-class prediction problems. For multi-class modeling, SVM-based compound ranking schemes,[15,16] rather than pairwise binary classifications, are particularly suitable. In general, multi-class predictions,[15,17] require the combination or sequential consideration of different SVM models. As an approach for model combination, SVM linear combination has previously been introduced,[18] which was originally applied to the prediction of ligands for orphan targets,[18] another chemogenomics application.

In our current study, we have investigated SVM modeling for another multi-class prediction problem involving compounds with overlapping activities against related targets. This application was found to be challenging for SVM ranking. Furthermore, a standard (unweighted) linear combination was not applicable in this case. Given the difficulties observed in distinguishing between compounds with in part overlapping activities using individual SVM models, we have combined independently derived models in differentially weighted SVM linear combination using positive and negative factors. The application of this strategy led to the preferential detection of compounds with desired activity profiles.

MATERIALS AND METHODS

Basic SVM Theory. As a supervised machine learning technique, SVMs[5] are primarily used for binary object classification and ranking. For learning, "positive" and "negative" training data (e.g., active and inactive compounds) are projected into a feature (descriptor) space χ. By solving a convex quadratic optimization problem, a hyperplane H is derived that best separates objects with different class labels. During the optimization, the trade-off parameter C is adjusted to balance errors due to misclassification of training data and the generalization of the classification. The separating hyperplane H is defined by the normal weight vector $\mathbf{w}$ and a bias b

$$H = \{\mathbf{x} | \langle \mathbf{w}, \mathbf{x} \rangle + b = 0\}, \text{ with } \langle \cdot, \cdot \rangle \text{ being a scalar product}$$

Test data are mapped into the same feature space χ and classified depending on which side of the generated hyperplane they fall. For SVM-based ranking compounds are sorted on the basis of the signed distance from the hyperplane (from the positive to the negative half-space):[16] $g(\mathbf{x}) = \langle \mathbf{w},\mathbf{x} \rangle$.

For training data that are nonlinearly separable in the feature space χ, which is usually the case, the so-called Kernel trick[19,20] is applied. For this purpose, kernel functions are utilized that replace an

Received: February 5, 2013
Published: March 21, 2013

dx.doi.org/10.1021/ci400090t | J. Chem. Inf. Model. 2013, 53, 791−801

explicit projection of the data into a high-dimensional feature space H where linear separation might be possible. Therefore, kernel functions $K(\cdot,\cdot)$ replace the standard scalar product.

Weighted SVM Linear Combination. The SVM linear combination (LC) was introduced as an extension of standard SVM classification to enable the prediction of ligands that are active against different targets.[18] In SVM LC, a hyperplane is generated for each individual target t_i under consideration using known active compounds of t_i as positive and inactive compounds as negative training data. The normal vectors of all hyperplanes are then linearly combined into a single vector

$$\mathbf{w}_{\text{combined}} = \sum_{i=1}^{n} f_i \mathbf{w}_i$$

where $\mathbf{w}_{\text{combined}}$ is the single combined normal vector, n the number of original hyperplanes, and f_i and $\mathbf{w}_i$ are the individual linear factors and normal vectors of each hyperplane, respectively. Herein, the LC approach is extended through the use of positive and negative linear factors. Test compounds are then ranked using a global ranking function

$$g(\mathbf{x}) = K(\mathbf{w}_{\text{combined}}, \mathbf{x})$$

Factors can be applied to adjust the relative contributions of individual weight vectors to the linear combination. Such weighting factors were previously applied to an SVM linear combination in similarity search calculations taking potency information of reference compounds into account.[21] In this case, SVM models derived for reference compounds at different potency levels were linearly combined, and their potency values were used as weighting factors.[21]

For the multi-class prediction task addressed in our current analysis, we introduce a modification of the weighted SVM LC approach that uses negative linear factors, as illustrated in Figure 1. The use of

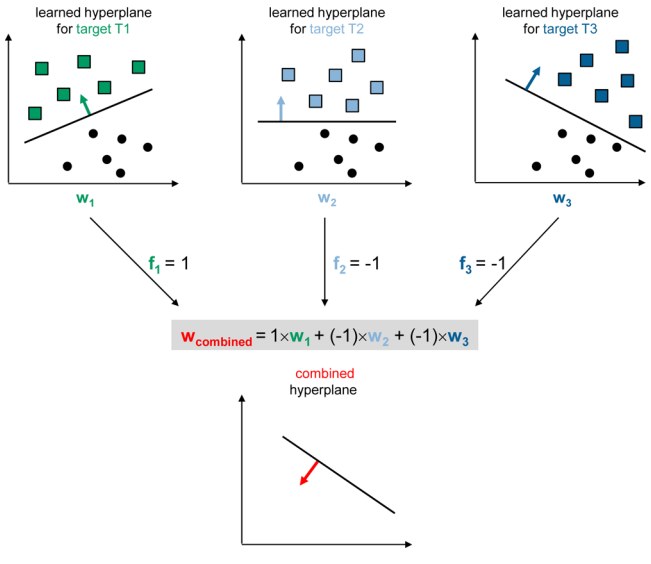


Figure 1. Weighted SVM linear combination. Normal weight vectors are derived from SVM calculations for sets of compounds active against three different targets. Weight vectors are linearly combined to yield a single vector used for global ranking. Positive and negative factors are applied to individual weight vectors to adjust their relative contributions to class label prediction or ranking.

negative factors effectively inverts the contributions of positive and negative training data for individual models during linear combination, a strategy that has not yet been considered in SVM modeling. This modification deprioritizes compounds with undesired activity relative to confirmed inactive compounds and compounds with desired activity. Balancing SVM LC through the application of positive as well as negative weighting factors is shown herein to effectively distinguish between compounds with different activity profiles.

In order to consistently prioritize or deprioritize compounds with given activity profiles, weighting factors of 1, 2, −1, and −2 were systematically varied. Single-target SVM classifiers were built as a control for each individual activity.

Compound Data Sets. Compound data sets with single-, dual-, and triple-target activities have been assembled from PubChem[22] confirmatory bioassays for three cytochrome P450 isoforms and three different dehydrogenases. The first set of three assays identified compounds active against cytochrome P450 2C19 (CYP2C19; assay id (AID) 899), cytochrome P450 2D6 (CYP2D6; AID 891), and cytochrome P450 3A4 (CYP3A4; AID 884). The second set of three assays contained inhibitors of aldehyde dehydrogenase 1 (ALDH1A1; AID 1030), hydroxyacyl-coenzyme A dehydrogenase type II (HADH2; AID 886), and 15-hydroxyprostaglandin dehydrogenase (HPGD; AID 894). From all assays confirmed active and inactive compounds were extracted and compared. Compounds with confirmed single-, dual-, and triple-target annotations were identified as well as compounds inactive against all targets. For cytochrome P450s, a total of 1906 active compounds with different target profiles were obtained and 2901 inactive compounds. For the dehydrogenases, 5085 active and 39,355 inactive compounds were obtained. The composition of the two data sets is reported in Figure 2.

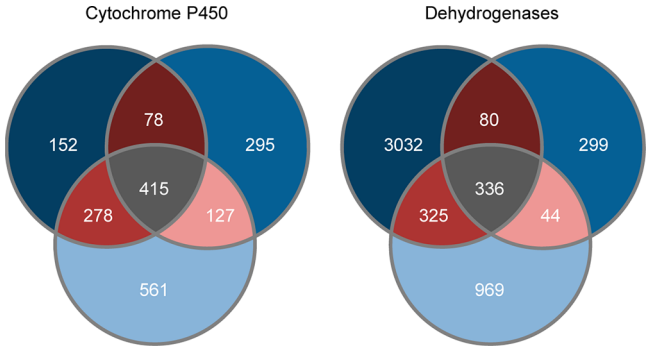


Figure 2. Active compounds. For inhibitors of three cytochrome P450 isoforms (data set 1) and three different dehydrogenases (data set 2), the number of compounds with single and multiple target annotations is reported in a Venn diagram.

Calculations. Compounds were represented using the extended-connectivity fingerprint[23] with bond diameter 4 (ECFP4) calculated using the Molecular Operating Environment.[24] For compound comparison during SVM calculations, the Tanimoto kernel[20] was applied.

In calculations searching for compounds with single-target annotations, compounds with the desired activity were used as positive training data and confirmed inactive compounds as negative data. In calculations searching for compounds with dual-target annotations, compounds with the desired dual-target activity were used as positive training data and confirmed inactive compounds as negative data. To assess search performance, SVM ranking was carried out using target-specific models and their weighted linear combination.

For training, 500 inactive compounds were used in each case as inactive training data. In order to use size-balanced compound sets for the generation of hyperplanes for linear combination, the number of

dx.doi.org/10.1021/ci400090t | J. Chem. Inf. Model. 2013, 53, 791−801

positive training compounds was set to half of the smallest available single- or dual-target compound subset for each series of search calculations.

For each search calculation, 100 different trials with randomly assembled positive and negative training data and test sets were carried out. In each case, test data consisted of all active compounds with different activity profiles plus all confirmed inactive compounds not used for training. Active and inactive compounds used for training were never included in test sets.

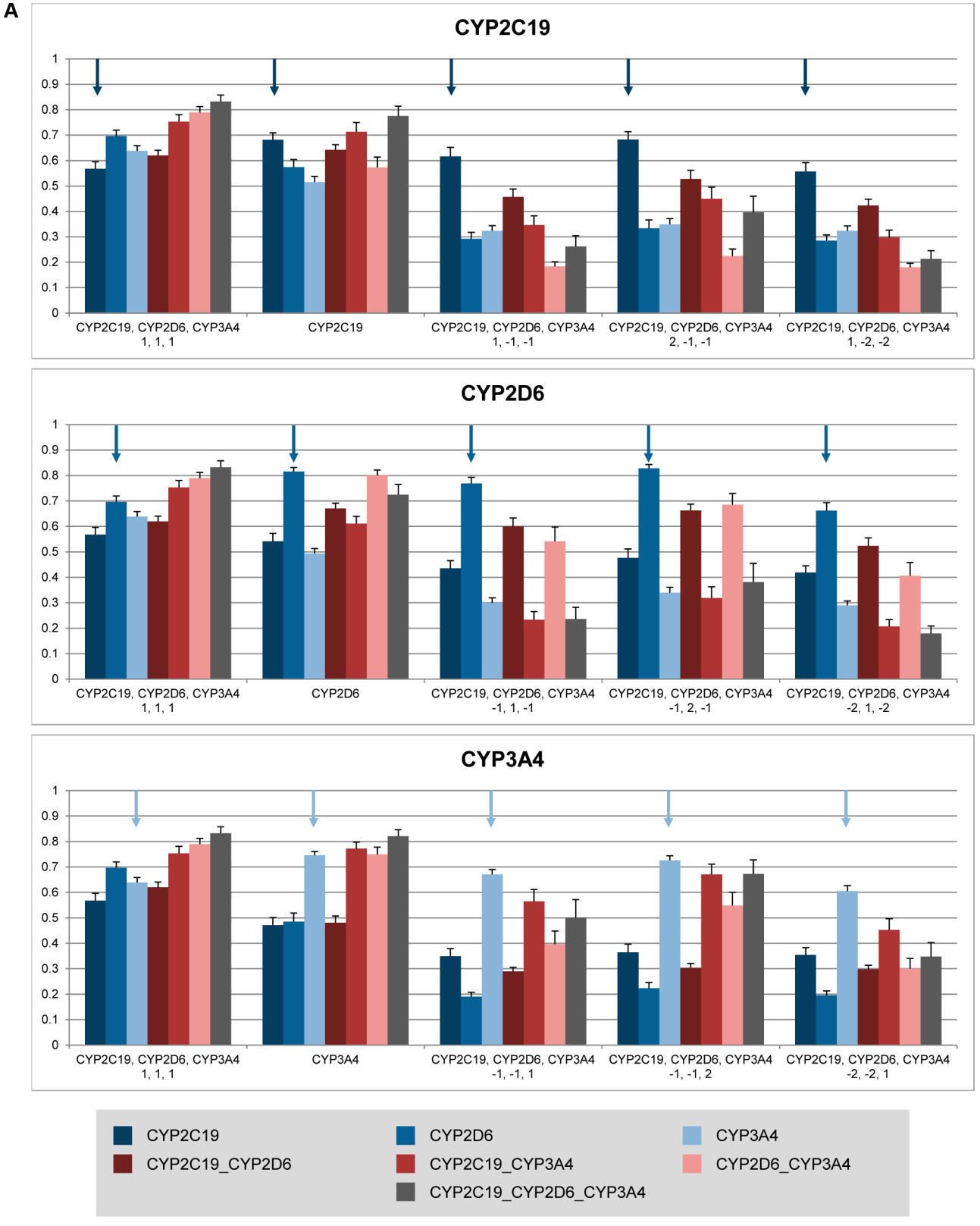


Figure 3. continued

dx.doi.org/10.1021/ci400090t | J. Chem. Inf. Model. 2013, 53, 791−801

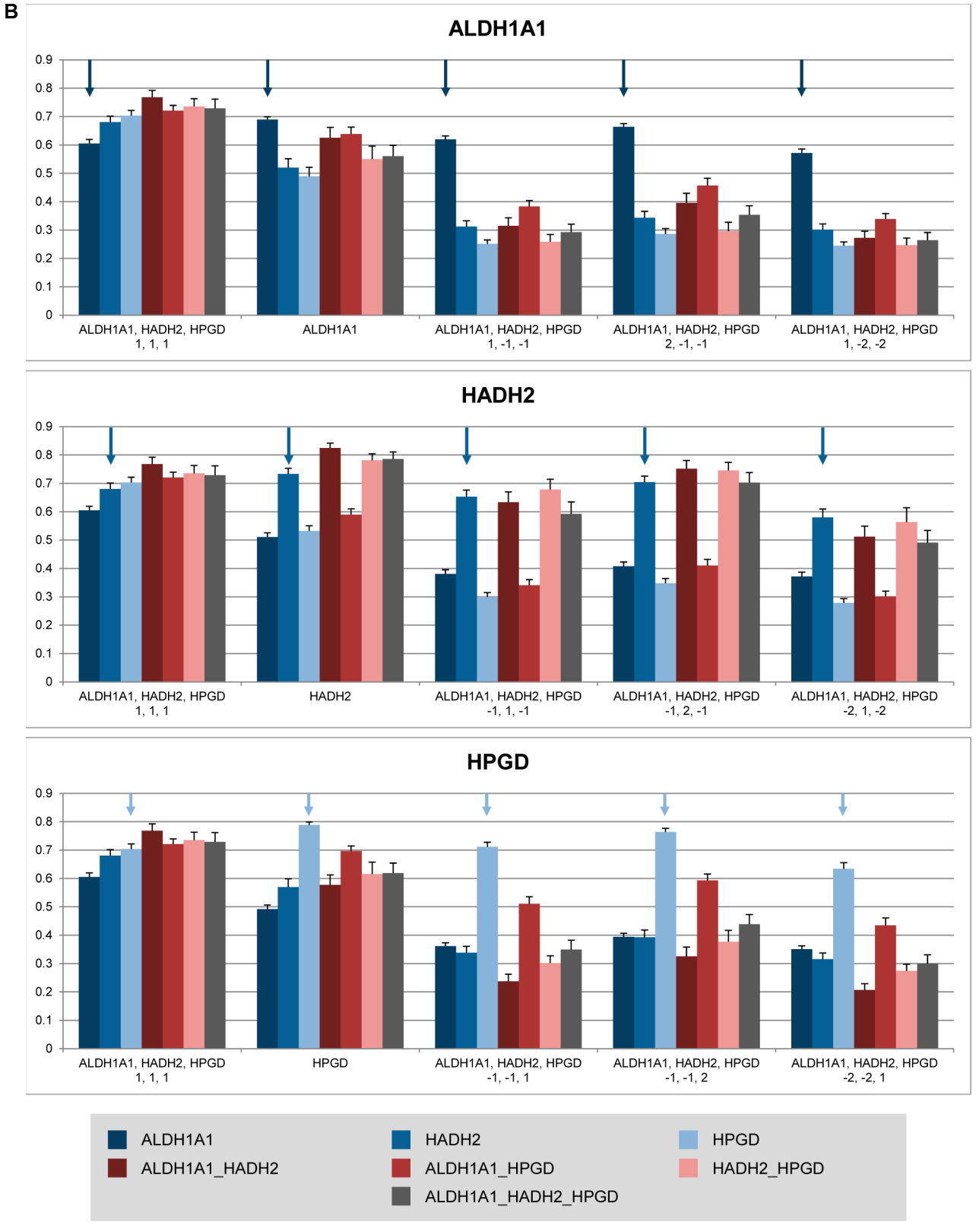


Figure 3. Searching compounds with single-target activities. ROC_AUC values are reported for compounds active against all possible target combinations in search calculations for compounds with single-target activities. (A) cytochrome P450s, (B) dehydrogenases. Dual- and triple-target combinations are indicated by underscores (e.g., CYP2C19_CYP2D6). Individual targets and target combinations are color-coded as in Figure 2. Search results are reported for standard SVM LC, SVM training using compounds active against the designated target, and differently weighted SVM LCs. In each case, color-coded arrows indicate the desired search result. In addition, standard deviations over 100 independent search trials are reported above each bar.

dx.doi.org/10.1021/ci400090t | J. Chem. Inf. Model. 2013, 53, 791−801

Search performance was measured using the area under the receiver operating characteristic curve (ROC_AUC)[25] averaged over all 100 trials.

SVM calculations were performed using SVM^{light},[26] a freely available SVM implementation. SVM parameters were suggested SVM^{light} default settings. SVM LC calculations were carried out using in-house generated Perl scripts.

RESULTS AND DISCUSSION

Compounds with Overlapping Activity Profiles. The compound data sets analyzed herein were assembled to investigate a multi-class prediction problem that we considered rather challenging: differentiating between compounds with related and overlapping activity profiles. The cytochrome P450 (CYP) and dehydrogenase data sets contained compounds with all possible combinations of single-, dual-, and triple-target activities, as illustrated in Figure 2. Because many compounds shared activities against individual targets in different combinations, this classification problem was anticipated to be difficult to solve using conventional SVM strategies.

Confirmed Inactive Compounds. With these data sets, we also addressed a potential caveat for machine learning in chemoinformatics that is often pointed out: usually randomly chosen sets of database compounds assumed to be inactive are used as negative training examples. By contrast, by assembling our data sets from PubChem confirmatory bioassays, we were able to obtain sets of confirmed inactive compounds for training against all possible activity combinations, hence providing a sound basis for model building. Our data sets containing all active and inactive compounds are made freely available via http://www.lifescienceinformatics.uni-bonn.de/downloads.

Single-Target Classifiers vs SVM Linear Combination. It should be stressed that SVM models built for individual activities, i.e., single-target classifiers, are conceptually distinct from SVM LC models. An SVM LC represents a model for multi-class predictions that integrates individual classifiers (and is as such distinct from them) and weights their relative contributions prior to predicting test data (but not posthoc). Hence, there is no retrospective fitting of individual classifiers comprising an SVM LC. Weighting factors and their combinations must be systematically explored, as discussed in the following.

Searching for Compounds with Single-Target Activity. We first generated SVM models on the basis of reference compounds that were active against individual targets to search for compounds with single-target activity. Using these models, compounds with all activity combinations were ranked and ROC_AUC values calculated for all compound categories. The results for compounds active against CYP isoforms and dehydrogenases are reported in Figure 3A andB (second bar charts from the left), respectively. For all three CYP targets, the recall of compounds with desired single-target activity was met or exceeded by compounds with activity against other targets or target combinations. Equivalent observations were made for dehydrogenase target HADH2, while for targets ALDH1A1 and HPGD at least slightly higher recall of specifically active compounds was observed. Overall, however, single-target SVM models failed to yield a clear separation in recall between compounds with the desired single-target activity and other activity profiles. Figure 3 also shows that single-target models produced comparably high recall of active compounds in most cases, with ROC_AUC values of ~0.7 to ~0.8, but failed to discriminate between compounds with different activity profiles. In addition, the standard SVM LC of single-target models (with factor setting "1,1,1" in Figure 3) preferentially detected compounds with triple- or dual-target activity. Differences in recall performance were small in all cases.

On the basis of these findings, we then investigated combinations of negative and positive factors for SVM linear combination, as rationalized in the Methods section. We first used a factor setting of "1,−1,−1" to deprioritize compounds with undesired single-target activities compared to the desired activity. Because undesired single-target activities were a part of all activity combinations, this factor setting was also anticipated to deprioritize compounds with dual- and triple-target activity. The results of weighted SVM LC calculations using positive and negative factors are reported in Figure 3 and confirmed the utility of this strategy. For all CYP targets, the recall of compounds with correct target activity was retained or only slightly reduced, whereas the recall of all other categories of compounds was significantly reduced, frequently to ROC_AUC values close to or below 0.5 (corresponding to random selection), as shown in Figure 3A. For compounds active against two dehydrogenases (ALDH1A1 and HPGD), equivalent observations were made. By contrast, in the case of HADH2, the recall of two dual- and the triple-target combinations remained comparable to compounds with the desired single-target activity, and no separation was observed (Figure 3B).

We then tested additional weighting factor combinations. For factor setting "2,−1,−1", which put additional weight on positive training examples with desired activity, a further increase in recall was generally observed for the desired compounds as expected, while the recall rates for other compound categories essentially remained constants for two of three CYP and two of three dehydrogenase targets. In the remaining two cases (CYP3A4 and HADH2), the recall for compounds with dual- and triple-target activities increased relative to the recall of desired compounds, which notably reduced the separation.

In addition, for factor setting "1,−2,−2", which more strongly deprioritized compounds with undesired activity, a significant reduction of recall of compounds with desired single-target activity (and in part other activity combinations) was observed, very likely because too much weight was put on negative training examples, hence rendering these calculations overall less sensitive to active compounds.

Among the differently weighted SVM LCs including negative factors, the factor settings "1,−1,−1" and "2,−1,−1" yielded a clear separation in recall between compounds with desired single-target activity and compounds with other activity profiles for four of six targets, whereas standard SVM calculations consistently failed to do so, despite reaching overall high recall performance on active compounds.

For calculations using SVM models trained on compounds with single-target activity, the recall of compounds with desired single-target activity varied between 0.68 for CYP2C19 and 0.82 for target CYP2D6. For standard SVM LC, the recall of compounds with desired single-target activity varied between 0.57 for CYP2C19 and 0.7 for HPGD. In differentially weighted SVM LCs, recall of the single-target activities ranged from 0.62 for CYP2C19 to 0.77 for CYP2D6.

As a consequence of SVM LC weighting, recall of undesired targets and target combination was reduced to values between 0.18 (CYP2D6_CYP3A4 compounds in single-target CYP2C19 calculation using factor "1,−1,−1") and 0.69 (CYP2D6_CYP3A4 compounds in single-target CYP2D6 calculation with factor "2,−1,−1"), with the majority of compound categories falling to ROC_AUC value below 0.5 (random selection).

Searching for Compounds with Dual-Target Activity. We next searched for compounds with dual-target activity. In these calculations, compounds with desired dual-activity were used as reference compounds for the generation of individual SVM models or SVM LCs. The linear combination strategy was adjusted accordingly. In this case, only two models were combined including the SVM model trained on the basis of the desired dual-target activity and the model derived for compounds with single-target activity against the third (undesired) target. This linear combination scheme enabled a meaningful application of

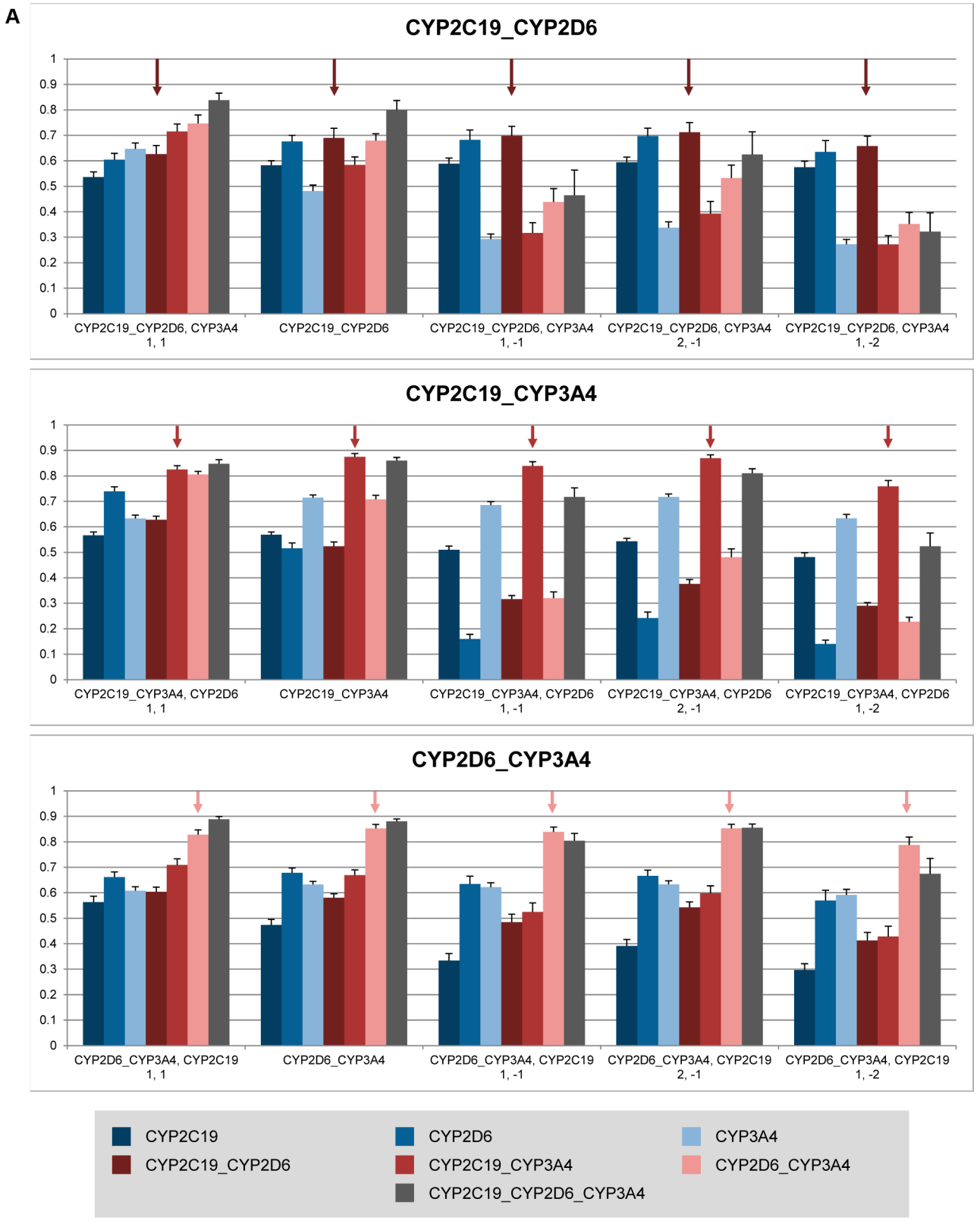


Figure 4. continued

dx.doi.org/10.1021/ci400090t | J. Chem. Inf. Model. 2013, 53, 791−801

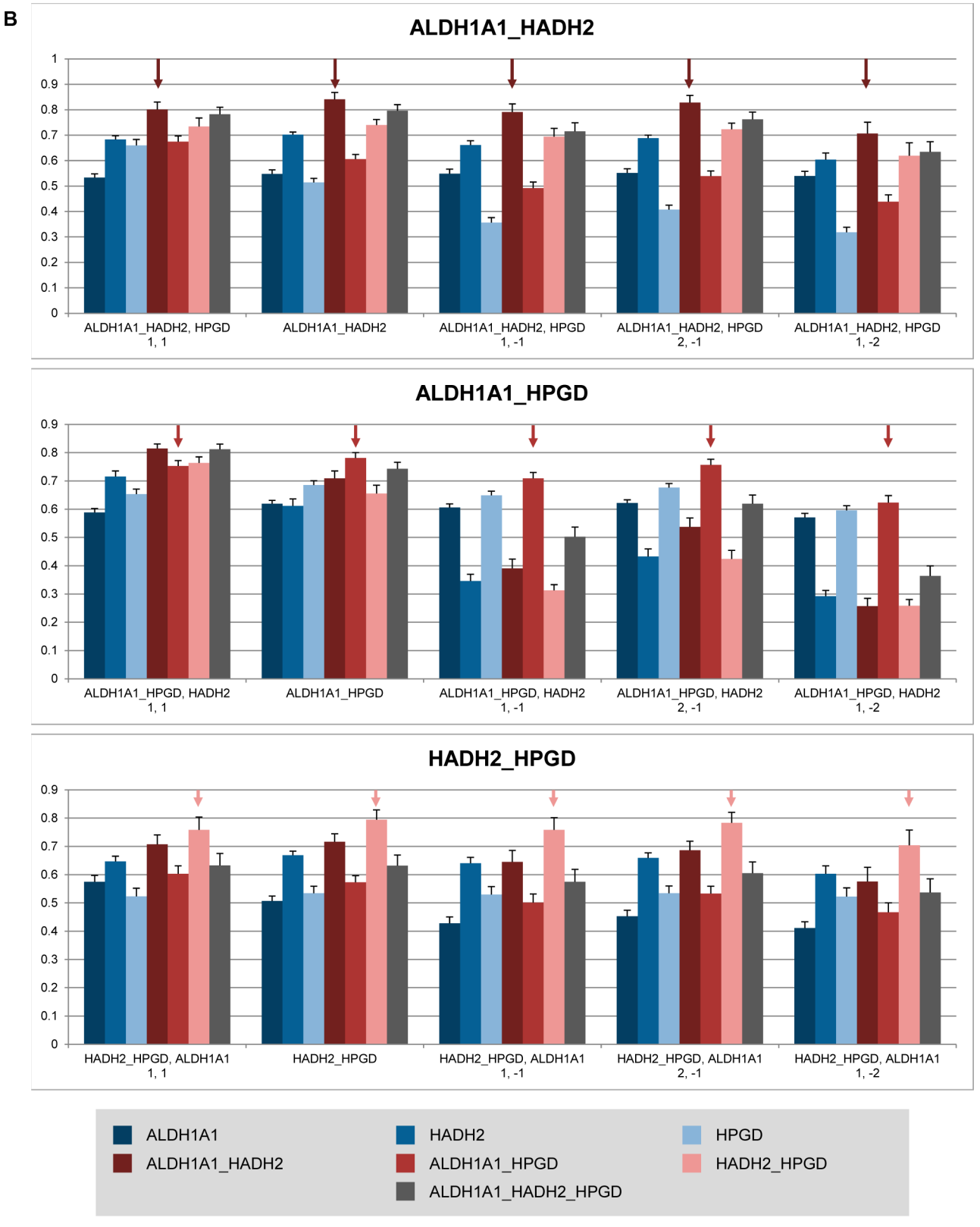


Figure 4. Searching compounds with dual-target activities. ROC_AUC values are reported for compounds active against all possible target combinations in search calculations for compounds with dual-target activities. (A) cytochrome P450s, (B) dehydrogenases. The presentation is according to Figure 3. Search results are reported for standard SVM LC, SVM training using compounds active against the designated target combination, and differently weighted SVM LCs.

negative and positive weighting factors. The search results for CYP isoforms and dehydrogenases are reported in Figure 4A and B, respectively.

Results obtained for SVM models generated for compounds with the desired dual-target activity and the standard SVM LC essentially paralleled the observations discussed above.

dx.doi.org/10.1021/ci400090t | J. Chem. Inf. Model. 2013, 53, 791−801

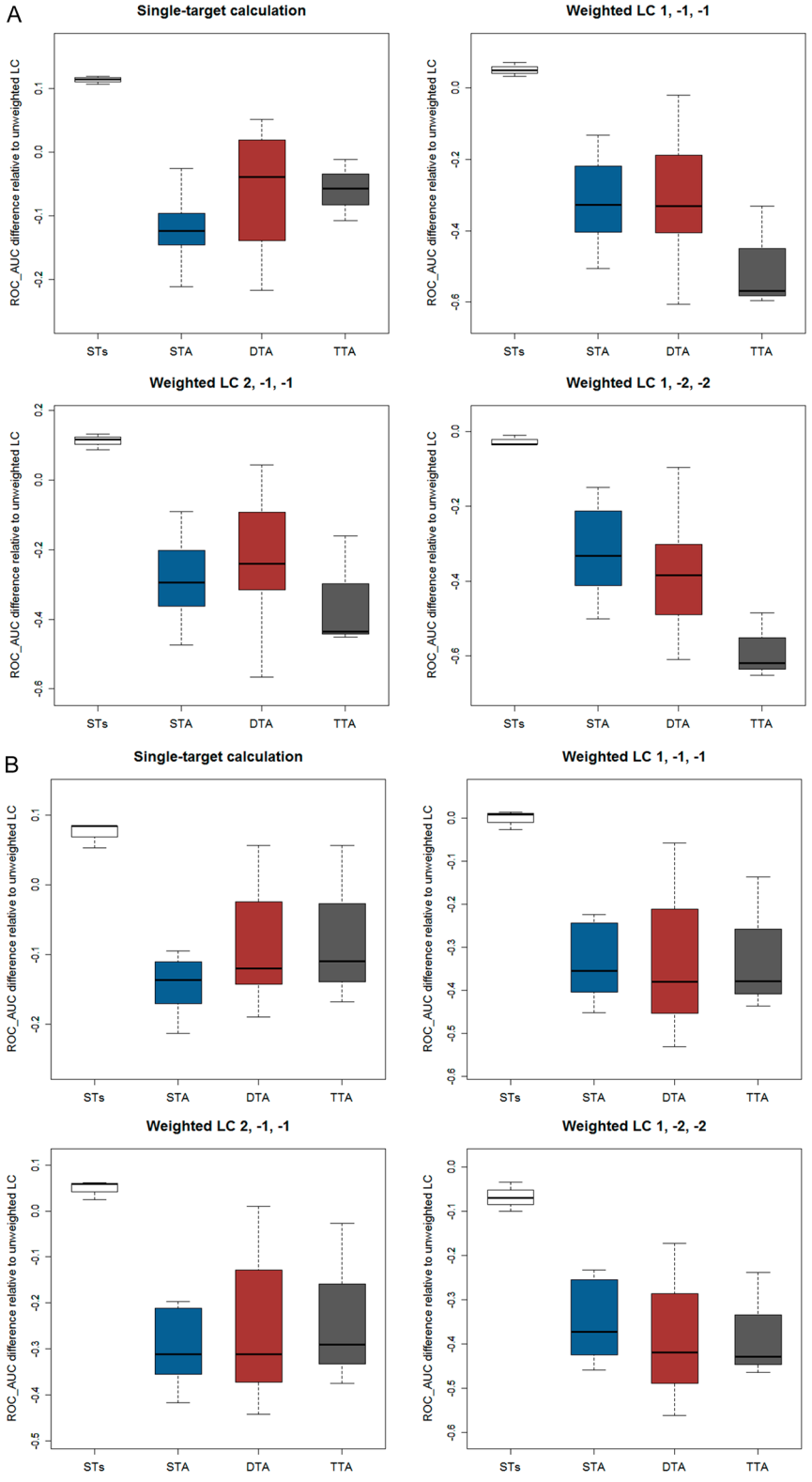


Figure 5. Recall differences for compounds with single-target activities. Differences between ROC_AUC values are reported for single-target calculations (Figure 3) and differently weighted SVM LCs relative to the standard (unweighted) LC calculation. (A) cytochrome P450s, (B) dehydrogenases. Boxplots are shown for four different types of targets or target combinations: desired single targets (STs, i.e., monitoring calculations for all three individual targets), other single-target activities (STA; blue), dual-target activities (DTA; red), and triple-target activity (TTA; gray).

The recall was high for many compound categories and no notable separation was observed. In addition, in this case, recall was high for compounds with triple-target activity in most calculations, as one would expect. Thus, standard SVM calculations also failed here to distinguish between compounds with different activity profiles.

When SVM LCs were tested with factor settings of "1,−1", which deprioritized compounds with activity against the undesired target,

dx.doi.org/10.1021/ci400090t | J. Chem. Inf. Model. 2013, 53, 791−801

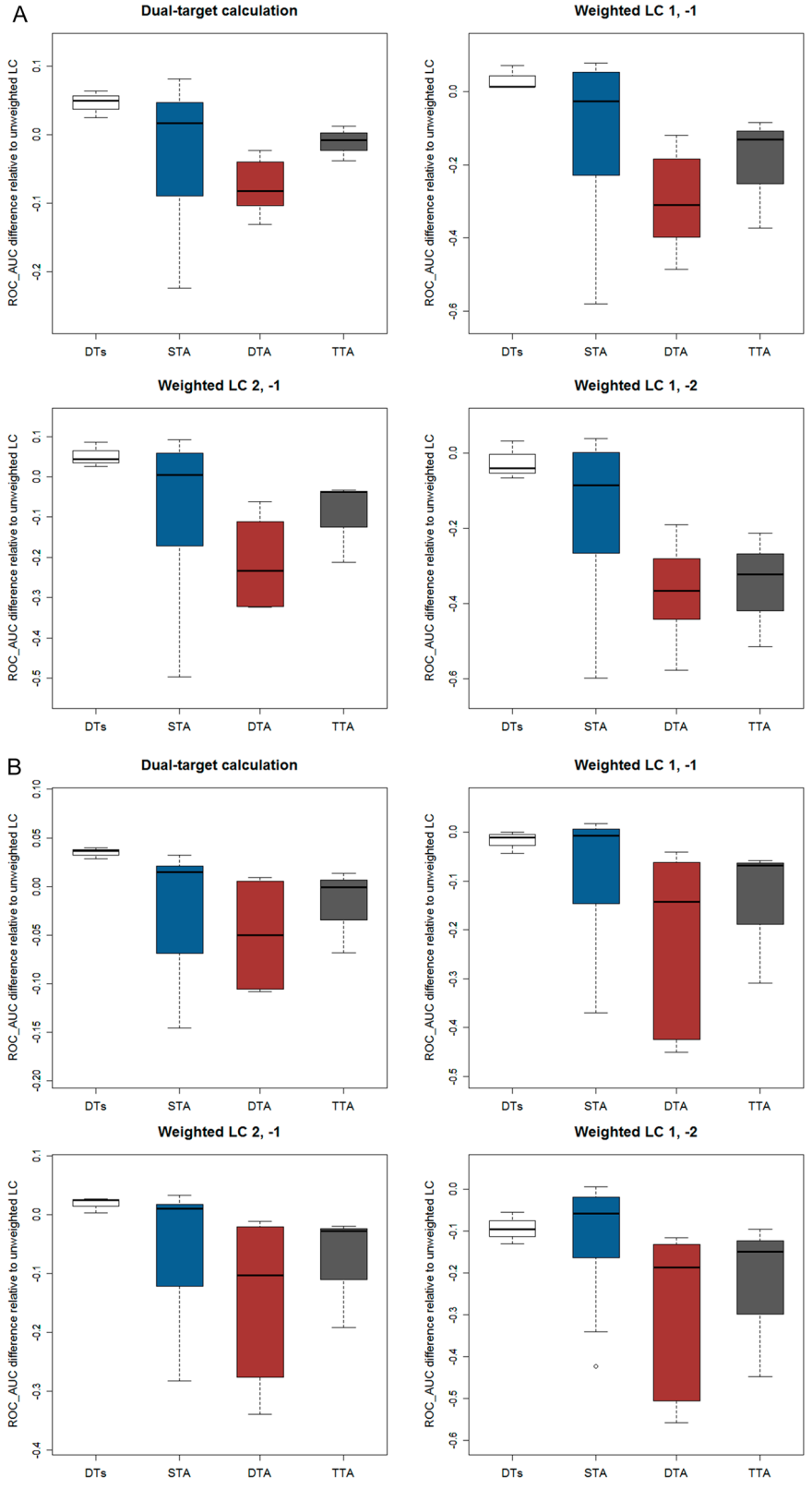


Figure 6. Recall differences for compounds with dual-target activities. Differences between ROC_AUC values are reported for dual-target calculations (Figure 4) and differently weighted SVM LCs relative to the standard LC calculation. (A) cytochrome P450s, (B) dehydrogenases. Boxplots are shown for four different types of targets or target combinations: desired dual-target combination (DTs, i.e., all calculations for the desired dual target combinations), STA, DTA, and TTA. Abbreviations and colors are used according to Figure 5.

and with setting "2,−1", which prioritized compounds with desired dual-target activity and deprioritized compounds with activity against the undesired target, a comparable improvement in recall separation was observed for four of six target combinations, except CYP2D6_CYP3A4 (Figure 4A) and HADH2_HPGD (Figure 4B). However, recall of compounds with activity against individual targets of the

dx.doi.org/10.1021/ci400090t | J. Chem. Inf. Model. 2013, 53, 791−801

desired combination and/or the triple-target combination remained high in most instances. Hence, a partial separation was observed in these cases, different from the calculations focusing on single-target activities discussed above. This was not unexpected given the dual-target nature of the desired activity profiles.

Furthermore, the application of factor setting "1,−2", which strongly deprioritized compounds active against the undesired target, led to a general reduction in recall, similar to the findings discussed for single-target activities under equivalent SVM LC weighting conditions because the influence of negative training examples was further emphasized in these cases.

Taken together, the results of search calculations obtained for compounds with dual-target activities indicated that deprioritization of the undesired target using weighting factor "−1" led to a preferred recall separation.

For standard SVM LC, the recall of compounds with desired dual-target activity varied between 0.63 for the CYP2C19_CYP2D6 and 0.83 for the CYP2C19_CYP3A4 combination. Because of SVM LC weighting, recall of undesired targets or target combinations including the undesired target was reduced to values between 0.16 (CYP2D6 compounds in calculations for CYP2C19_CYP3A4 using factor "1,−1") and 0.86 (for the triple-target set of all CYP isoforms in calculations for CYP2D6_CYP3A4 with factor "2,−1").

Recall Differences. The recall trends and separation effects discussed above are further quantified for single-target and dual-target calculations in Figures 5 and 6, respectively. As a reference point for all comparisons, standard SVM LC recall was used. In search calculations for compounds with single-target activities, median ROC_AUC value differences in recall between compounds with desired activity and other activity profiles were within 0.2 for SVM models trained on compounds with single target activity for both CYP (Figure 5A) and dehydrogenase targets (Figure 5B). In both cases, SVM LCs with negative weighting factors increased the recall separation to median ROC_AUC value differences of ∼0.3 to ∼0.6, depending on the model and compound category. In search calculations for compounds with dual-target activities, maximal median ROC_AUC value separations of ∼0.15 and ∼0.1 were observed for CYP (Figure 6A) and dehydrogenase targets (Figure 6B), respectively, when SVM models trained on compounds with dual-target activities were utilized. For weighted SVM LCs, the median recall for compounds with single-target activity was very similar to recall for compounds with activity against the desired target combination (due to the influence of shared targets), but the separation relative to compounds with other dual-target or triple-target activity was increased to median values of ∼0.2 to ∼0.4 for CYP and ∼0.05 to ∼0.2 for dehydrogenase targets. For individual compound sets, much larger recall separations were also observed in the latter case, as shown in Figure 6B.

CONCLUSIONS

In this study, we have investigated a multi-class prediction task involving compounds with activity against different combinations of targets. Given the overlap in activity profiles between these compounds, we anticipated that it might be difficult to address this task. Initially, individual SVM models were trained for all compound categories. Searching for compounds with desired single- or dual-target activity in the presence of confirmed inactive compounds using standard SVM calculations confirmed our expectations. SVM-based compound ranking was found to produce reasonable to high compound recall for different compound categories but essentially failed to distinguish compounds with desired activity from compounds with other activity profiles. Therefore, we designed an SVM linear combination strategy that involved weighting of different models using positive and negative factors. The combination of models and use of positive and negative weighting factors made it possible to prioritize and deprioritize compounds with desired and undesired activity profiles, respectively. Differentially weighted SVM LC calculations yielded in part significant recall separation effects. Especially for compounds with desired single-target activity, the weighted SVM LC approach consistently reduced the recall of compounds with different activity profiles while essentially retaining the recall of compounds with desired target activity. Hence, the SVM LC weighting strategy introduced herein to investigate a complex activity prediction task should also be of interest for other multi-class SVM applications.

AUTHOR INFORMATION

Corresponding Author

*Tel: +49-228-2699-306. Fax: +49-228-2699-341. E-mail: bajorath@bit.uni-bonn.de.

Notes

The authors declare no competing financial interest.

REFERENCES

(1) Duda, R. O.; Hart, P. E.; Stork, D. G. *Pattern Classification*, 2nd ed.; Wiley-Interscience: New York, 2000, pp 20−83.

(2) Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S. Molecular similarity searching using atom environments, information-based feature selection, and a naïve Bayesian classifier. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 170−178.

(3) Nidhi; Glick, M.; Davies, J. W.; Jenkins, J. L. Prediction of biological targets for compounds using multiple-category Bayesian models trained on chemogenomics databases. *J. Chem. Inf. Model.* **2006**, *46*, 1124−1133.

(4) Watson, P. Naive Bayes classification using 2D pharmacophore feature triplet vectors. *J. Chem. Inf. Model.* **2008**, *48*, 166−178.

(5) Vapnik, V. N. *The Nature of Statistical Learning Theory*, 2nd ed.; Springer: New York, 2000.

(6) Burges, C. J. C. A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Discovery* **1998**, *2*, 121−167.

(7) Burbidge, R.; Trotter, M.; Holden, S.; Buxton, B. Drug design by machine learning: Support vector machines for pharmaceutical data analysis. *Comput. Chem.* **2001**, *26*, 5−14.

(8) Warmuth, M. K.; Liao, J.; Rätsch, G.; Mathieson, M.; Putta, S.; Lemmen, C. Active learning with support vector machines in the drug discovery process. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 667−673.

(9) Jorissen, R. N.; Gilson, M. K. Virtual screening of molecular databases using a support vector machine. *J. Chem. Inf. Model.* **2005**, *45*, 549−561.

(10) Bredel, M.; Jacoby, E. Chemogenomics: An emerging strategy for rapid target and drug discovery. *Nat. Rev. Genet.* **2004**, *5*, 262−275.

(11) Stockwell, B. R. Exploring biology with small organic molecules. *Nature* **2004**, *432*, 846−854.

(12) Paolini, G. V.; Shapland, R. H. B.; van Hoorn, W. P.; Mason, J. S.; Hopkins, A. L. Global mapping of pharmacological space. *Nat. Biotechnol.* **2006**, *24*, 805−815.

(13) Bock, J. R.; Gough, D. A. Virtual screens for ligands of orphan G protein-coupled receptors. *J. Chem. Inf. Model.* **2005**, *45*, 1402−1414.

(14) Jacob, L.; Vert, J.-P. Protein−ligand interaction prediction: An improved chemogenomics approach. *Bioinformatics* **2008**, *24*, 2149−2156.

(15) Wassermann, A. M.; Geppert, H.; Bajorath, J. Searching for target-selective compounds using different combinations of multi-class support vector machine ranking methods, kernel functions, and fingerprint descriptors. *J. Chem. Inf. Model.* **2009**, *49*, 582−592.

(16) Geppert, H.; Horváth, T.; Gärtner, T.; Wrobel, S.; Bajorath, J. Support-vector-machine-based ranking significantly improves the

effectiveness of similarity searching using 2D fingerprints and multiple reference compounds. *J. Chem. Inf. Model.* **2008**, *48*, 742−746.
(17) Kawai, K.; Fujishima, S.; Takahashi, Y. Predictive activity profiling of drugs by topological-fragment-spectra-based support vector machines. *J. Chem. Inf. Model.* **2008**, *48*, 1152−1160.
(18) Geppert, H.; Humrich, J.; Stumpfe, D.; Gärtner, T.; Bajorath, J. Ligand prediction from protein sequence and small molecule information using support vector machines and fingerprint descriptors. *J. Chem. Inf. Model.* **2009**, *49*, 767−779.
(19) Boser, B. E.; Guyon, I. M.; Vapnik, V. N. A Training Algorithm for Optimal Margin Classifiers. In *Proceedings of the 5th Annual Workshop on Computational Learning Theory*; Pittsburgh, PA, 1992, ACM: New York, 1992; pp 144−152.
(20) Ralaivola, L.; Swamidass, S. J.; Saigo, H.; Baldi, P. Graph kernels for chemical informatics. *Neural Networks* **2005**, *18*, 1093−1110.
(21) Wassermann, A. M.; Heikamp, K.; Bajorath, J. Potency-directed similarity searching using support vector machines. *Chem. Biol. Drug Des.* **2011**, *77*, 30−38.
(22) Wang, Y.; Xiao, J.; Suzek, T. O.; Zhang, J.; Wang, J.; Zhou, Z.; Han, L.; Karapetyan, K.; Dracheva, S.; Shoemaker, B. A.; Bolton, E.; Gindulyte, A.; Bryant, S. H. PubChem's bioassay database. *Nucleic Acids Res.* **2012**, *40*, D400−D412.
(23) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.
(24) *Molecular Operating Environment (MOE)*; Chemical Computing Group, Inc.: Montreal, Quebec, Canada.
(25) Witten, I. H.; Frank, E. *Data Mining − Practical Machine Learning Tools and Techniques*, ed. 2; Morgan Kaufmann: San Francisco, 2005, pp 161−176.
(26) Joachims, T. Making Large-Scale SVM Learning Practical. In *Advances in Kernel Methods − Support Vector Learning*; Schölkopf, B., Burges, C. J. C., Smola, A. J., Eds.; MIT-Press: Cambridge, MA, 1999; pp 169−184.