JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

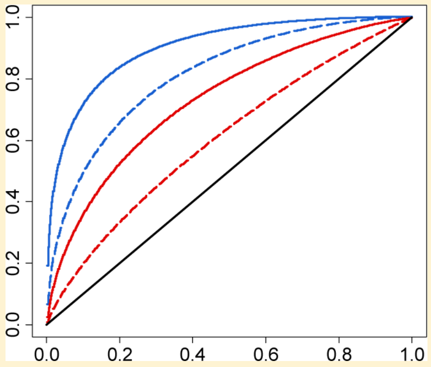
Comparison of Confirmed Inactive and Randomly Selected Compounds as Negative Training Examples in Support Vector Machine-Based Virtual Screening

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

Ⓢ *Supporting Information*

ABSTRACT: The choice of negative training data for machine learning is a little explored issue in chemoinformatics. In this study, the influence of alternative sets of negative training data and different background databases on support vector machine (SVM) modeling and virtual screening has been investigated. Target-directed SVM models have been derived on the basis of differently composed training sets containing confirmed inactive molecules or randomly selected database compounds as negative training instances. These models were then applied to search background databases consisting of biological screening data or randomly assembled compounds for available hits. Negative training data were found to systematically influence compound recall in virtual screening. In addition, different background databases had a strong influence on the search results. Our findings also indicated that typical benchmark settings lead to an overestimation of SVM-based virtual screening performance compared to search conditions that are more relevant for practical applications.



INTRODUCTION

For ligand-based virtual screening, selected machine-learning methodologies are increasingly utilized, given their usually good performance in predicting active compounds, at least in benchmark settings.[1,2] Methods like Bayesian modeling[3−6] and support vector machines (SVMs)[7−11] currently are among the most widely used methodologies for supervised learning to predict candidate compounds for different targets and to rank them according to their proposed likelihood of activity. These machine-learning algorithms rely on already known active compounds to derive computational models for compound classification and activity prediction.

The availability of sufficient amounts of relevant training data is a prerequisite for the applicability of these approaches. If no, only one, or very few known active compounds are available for a new target, scientifically sound classification models cannot be derived. For orphan screening, compound information from related targets (if available) must be utilized. Hence, knowledge of active compounds is generally considered the most critical requirement for model building. Consequently, major databases such as ChEMBL[12] and BindingDB,[13] which store active compounds from medicinal chemistry together with their activity data, are prime sources of positive training examples for machine learning.

A general caveat for training set assembly is that confirmed inactive compounds are often not available for a given target. Therefore, it is common practice in Bayesian or SVM modeling to randomly select compounds from databases that are not annotated with biological activities as negative training examples and assume that these random selections are inactive against the target of interest.[2,14−16] Scientifically, this represents an approximation, which is only very little investigated in chemoinformatics machine-learning applications.

The potential influence of negative training data on the quality of machine-learning models is just beginning to be addressed in the chemoinformatics field. In a recent study, Smusz et al.[17] tested alternative selection methods for assumed inactive training compounds. Using different fingerprints, machine-learning algorithms, and targets, the authors compared random and diverse selection of negative training examples from the ZINC database,[18] the Molecular Drug Data Report (MDDR),[19] and from compound libraries that were designed following the principles underlying the Directory of Useful Decoys (DUD) approach.[20] In these benchmark calculations, overall best compound classification results were achieved with different methods when negative training compounds were randomly selected from ZINC.[17] This database represents the largest public collection of compounds from chemical vendor sources. ZINC compounds, which are typically not biologically annotated, currently are the most popular source for random compound selection in machine-learning and ligand-based virtual screening.

Compounds confirmed to be inactive against a virtual screening target should generally have highest priority as negative training data for machine learning. This is the case

Received: May 8, 2013
Published: June 25, 2013

dx.doi.org/10.1021/ci4002712 | *J. Chem. Inf. Model.* 2013, 53, 1595−1601

because experimentally confirmed inactive compounds no longer rely on the assumption of inactivity that needs to be made for randomly chosen database compounds, which might or might not be true for individual molecules. Although it is difficult to obtain confirmed inactive compounds for many targets, the problem can be addressed by taking biological screening data into account. For example, confirmatory screening assays follow up on compounds with an activity signal in a primary screen (usually at a single concentration), investigate dose−response behavior, determine IC_{50} values for active compounds, and identify false positives. Thus, such follow-up assays confirm the activity of initial hits and also yield confirmed inactive compounds. Alternatively, dose−response behavior and IC_{50} titration curves might also be determined in a primary screen using multiple compound concentrations. In this case, large numbers of inactive compounds can be directly obtained.

In this study, we have investigated the influence of different negative training data and background databases on compound recall in SVM-based virtual screening. By assembling data sets from the PubChem Confirmatory Bioassays,[21] we were able to compare the search performance for training data sets comprising experimentally confirmed active and inactive compounds and training data sets consisting of experimentally confirmed active and randomly chosen "inactive" compounds. In addition, the size and relative composition of training data sets were varied and SVM-based virtual screening was carried out using either the biological screening database or ZINC compounds as a background. The results of these systematic calculations are reported in the following.

MATERIALS AND METHODS

Support Vector Machines. SVMs[7] are a supervised machine-learning technique for binary object classification and ranking. In the training phase, a set of "positive" and "negative" data are projected into a feature space χ. In ligand-based virtual screening, training objects are known active and assumed/known inactive compounds that are represented by a feature vector $\mathbf{x}_i \in \chi$. During optimization, a convex quadratic optimization problem is solved, and a hyperplane H is derived that best separates the positive and negative training objects from each other. Maximizing the margin (i.e., distance from the nearest training examples) and minimizing training errors are basic requirements to achieve model generalization and high prediction performance. The hyperplane H is defined by the normal weight vector $\mathbf{w}$ and a bias b, so that $H = \{\mathbf{x}|\langle\mathbf{w},\mathbf{x}\rangle + b = 0\}$, with $\langle\cdot,\cdot\rangle$ being a scalar product.

Test data, i.e., compounds with unknown activity, are also mapped into the feature space χ. Depending on which side of the hyperplane the test compounds fall, they are classified either as positive (active) or negative (inactive). In SVM ranking, compounds are sorted from the position most distant to the hyperplane on the positive half-space to the most distant position on the negative half-space using their score $g(\mathbf{x}) = \langle\mathbf{w},\mathbf{x}\rangle$.

In order to allow model building in the case of nonlinearly separable training data in the feature space χ, the so-called *Kernel trick*[22] is applied to replace the standard scalar product $\langle\cdot,\cdot\rangle$ by a kernel function $K(\cdot,\cdot)$. The kernel transfers the calculation of the scalar product into a higher dimensional space $\mathcal{H}$, where a linear separation might be feasible, without explicitly calculating the mapping into $\mathcal{H}$. This operation is at the core of SVM modeling.

As descriptors for SVM modeling, sets of numerical descriptors or fingerprints (i.e., bit string representations of molecular structure and properties) can be used.

Compound Data Sets. Six confirmatory high-throughput screening (HTS) assays have been extracted from PubChem, as reported in Table 1. These inhibitor assays were chosen to select targets from diverse families and maximize the number of confirmed active compounds available for modeling. From all assays, confirmed active and inactive compounds consisting of at least five non-hydrogen atoms were extracted. For each data set, a set of assumed inactive compounds was randomly selected from ZINC (version 12) that had the same size as the set of experimentally confirmed inactive compounds.

Table 1. Compound Data Sets[a]

AID	target	target code	# actives	# inactives
504332	euchromatic histone-lysine N-methyltransferase 2	EHMT2	30,170	262,493
1030	aldehyde dehydrogeanse 1	ALDH1A1	15,822	143,429
504333	bromodomain adjacent to zinc finger domain 2B	BAZ2B	15,539	307,386
504444	nuclear factor erythroid 2-related factor 2 isoform 2	Nrf2	7,284	280,339
588855	transforming growth factor beta	SMAD3	4,800	343,727
588591	polymerase eta	POLH	4,664	366,364

[a] Data sets were extracted from PubChem Confirmatory Bioassays. For each of the six data sets, the PubChem assay id (AID), the target name, and a target code are given. In addition, the numbers of confirmed active (# actives) and confirmed inactive compounds (# inactives) are reported. The data sets are sorted by decreasing numbers of active compounds.

Training Set Composition and Background Databases. SVM models were built using active training compounds from PubChem and either confirmed inactive PubChem compounds (P/P) or assumed inactive compounds randomly selected from ZINC (P/Z). In each case, training set sizes were varied to include all possible combinations of 100, 200, 500, or 1000 active and inactive training compounds for a total of 16 combinations with different proportions of active and inactive compounds. In each case, the remaining active compounds according to Table 1 were then added as potential hits to the background/screening database. Two background databases were explored in each case. The first database contained all remaining inactive compounds (not used for model building) from each PubChem screening set. Thus, the size of this background database varied in each case according to Table 1 but always contained hundreds of thousands of compounds. As the second background database, the same number of randomly selected ZINC compounds was used in each case. Combination of the two training set categories (P/P, P/Z) and the two background databases (P, Z) resulted in four distinct training/test categories designated as P/P−P, P/P-Z, P/Z-P, and P/Z-Z. These four training/test categories in combination with 16 different training set compositions then gave rise to 64 alternative screening setups.

Calculations and Performance Criteria. Compounds were represented using the extended-connectivity fingerprint[23] with bond diameter 4 (ECFP4) or MACCS structural keys[24] calculated with the Molecular Operating Environment.[25] For SVM model building, the Tanimoto kernel[26] was used. With

Table 2. Average AUC Values and Standard Deviations for ECFP4[a]

(A)

EHMT2	P/Z-Z		P/P−P		P/Z-P		P/P-Z	
# refs	AUC	SD	AUC	SD	AUC	SD	AUC	SD
100A_100I	0.800	0.011	0.647	0.017	0.641	0.013	0.692	0.042
100A_200I	0.820	0.008	0.666	0.014	0.661	0.009	0.706	0.026
100A_500I	0.840	0.003	0.679	0.006	0.671	0.006	0.708	0.015
100A_1000I	0.846	0.004	0.681	0.007	0.674	0.004	0.702	0.015
200A_100I	0.819	0.008	0.653	0.014	0.655	0.010	0.704	0.029
200A_200I	0.837	0.008	0.682	0.008	0.668	0.014	0.730	0.022
200A_500I	0.855	0.005	0.706	0.008	0.682	0.007	0.744	0.020
200A_1000I	0.865	0.004	0.711	0.009	0.686	0.005	0.740	0.016
500A_100I	0.833	0.005	0.676	0.013	0.664	0.009	0.721	0.032
500A_200I	0.851	0.006	0.702	0.007	0.678	0.010	0.750	0.013
500A_500I	0.873	0.004	0.728	0.005	0.695	0.006	0.776	0.010
500A_1000I	0.885	0.004	0.743	0.005	0.705	0.005	0.783	0.012
1000A_100I	0.837	0.006	0.679	0.011	0.663	0.009	0.721	0.023
1000A_200I	0.859	0.005	0.708	0.008	0.680	0.009	0.756	0.025
1000A_500I	0.882	0.002	0.739	0.004	0.702	0.006	0.794	0.018
1000A_1000I	0.896	0.002	0.761	0.002	0.713	0.004	0.809	0.008

(B)

ALDH1A1	P/Z-Z		P/P−P		P/Z-P		P/P-Z	
# refs	AUC	SD	AUC	SD	AUC	SD	AUC	SD
100A_100I	0.813	0.009	0.606	0.012	0.619	0.010	0.671	0.031
100A_200I	0.825	0.010	0.623	0.009	0.623	0.009	0.690	0.028
100A_500I	0.844	0.004	0.638	0.008	0.630	0.007	0.686	0.014
100A_1000I	0.847	0.008	0.646	0.011	0.628	0.009	0.689	0.021
200A_100I	0.830	0.008	0.624	0.010	0.631	0.009	0.687	0.027
200A_200I	0.843	0.010	0.645	0.010	0.635	0.010	0.709	0.016
200A_500I	0.863	0.008	0.669	0.011	0.644	0.007	0.722	0.019
200A_1000I	0.869	0.007	0.674	0.009	0.643	0.008	0.718	0.018
500A_100I	0.841	0.005	0.640	0.014	0.642	0.006	0.676	0.032
500A_200I	0.861	0.004	0.671	0.010	0.653	0.005	0.718	0.024
500A_500I	0.884	0.003	0.696	0.007	0.663	0.005	0.755	0.011
500A_1000I	0.893	0.004	0.712	0.005	0.666	0.005	0.760	0.016
1000A_100I	0.850	0.005	0.647	0.009	0.644	0.007	0.702	0.024
1000A_200I	0.872	0.006	0.678	0.008	0.654	0.006	0.730	0.020
1000A_500I	0.893	0.003	0.708	0.006	0.670	0.005	0.767	0.019
1000A_1000I	0.905	0.003	0.732	0.003	0.676	0.007	0.795	0.017

(C)

BAZ2B	P/Z-Z		P/P−P		P/Z-P		P/P-Z	
# refs	AUC	SD	AUC	SD	AUC	SD	AUC	SD
100A_100I	0.835	0.012	0.711	0.013	0.696	0.011	0.759	0.025
100A_200I	0.849	0.009	0.728	0.013	0.708	0.009	0.771	0.020
100A_500I	0.862	0.008	0.739	0.015	0.709	0.008	0.773	0.020
100A_1000I	0.868	0.007	0.747	0.010	0.708	0.008	0.771	0.015
200A_100I	0.852	0.008	0.724	0.011	0.710	0.006	0.771	0.021
200A_200I	0.867	0.009	0.747	0.011	0.728	0.007	0.790	0.016
200A_500I	0.880	0.007	0.765	0.005	0.732	0.007	0.798	0.009
200A_1000I	0.890	0.008	0.778	0.008	0.730	0.008	0.801	0.011
500A_100I	0.868	0.007	0.732	0.012	0.716	0.012	0.785	0.019
500A_200I	0.881	0.006	0.766	0.009	0.731	0.010	0.814	0.013
500A_500I	0.901	0.003	0.794	0.005	0.754	0.003	0.831	0.008
500A_1000I	0.911	0.004	0.810	0.006	0.759	0.005	0.837	0.009
1000A_100I	0.873	0.006	0.751	0.009	0.723	0.012	0.793	0.021
1000A_200I	0.892	0.004	0.782	0.008	0.741	0.008	0.826	0.014
1000A_500I	0.912	0.002	0.810	0.004	0.764	0.005	0.853	0.007
1000A_1000I	0.924	0.001	0.830	0.003	0.775	0.005	0.863	0.008

Table 2. continued

(D)

Nrf2	P/Z-Z		P/P–P		P/Z-P		P/P-Z	
# refs	AUC	SD	AUC	SD	AUC	SD	AUC	SD
100A_100I	0.776	0.005	0.639	0.013	0.614	0.008	0.666	0.029
100A_200I	0.781	0.008	0.653	0.011	0.609	0.011	0.666	0.020
100A_500I	0.795	0.011	0.669	0.009	0.607	0.013	0.674	0.017
100A_1000I	0.801	0.011	0.673	0.007	0.606	0.012	0.671	0.020
200A_100I	0.800	0.007	0.669	0.014	0.631	0.010	0.696	0.017
200A_200I	0.816	0.007	0.679	0.011	0.636	0.010	0.697	0.019
200A_500I	0.826	0.008	0.698	0.009	0.633	0.011	0.699	0.017
200A_1000I	0.837	0.008	0.706	0.007	0.629	0.009	0.703	0.017
500A_100I	0.817	0.008	0.689	0.012	0.650	0.010	0.705	0.013
500A_200I	0.833	0.007	0.714	0.009	0.658	0.011	0.726	0.011
500A_500I	0.856	0.005	0.740	0.007	0.671	0.007	0.744	0.017
500A_1000I	0.869	0.005	0.747	0.004	0.671	0.008	0.738	0.010
1000A_100I	0.824	0.006	0.696	0.007	0.656	0.011	0.708	0.014
1000A_200I	0.846	0.003	0.730	0.007	0.669	0.011	0.742	0.019
1000A_500I	0.868	0.002	0.760	0.008	0.684	0.006	0.770	0.010
1000A_1000I	0.885	0.004	0.771	0.003	0.692	0.006	0.770	0.006

(E)

SMAD3	P/Z-Z		P/P–P		P/Z-P		P/P-Z	
# refs	AUC	SD	AUC	SD	AUC	SD	AUC	SD
100A_100I	0.824	0.004	0.711	0.014	0.693	0.008	0.739	0.026
100A_200I	0.839	0.007	0.728	0.013	0.700	0.013	0.752	0.029
100A_500I	0.855	0.006	0.745	0.015	0.701	0.013	0.756	0.026
100A_1000I	0.859	0.006	0.750	0.009	0.701	0.012	0.754	0.018
200A_100I	0.834	0.008	0.734	0.013	0.704	0.008	0.767	0.018
200A_200I	0.856	0.004	0.753	0.008	0.723	0.007	0.787	0.017
200A_500I	0.872	0.003	0.773	0.006	0.725	0.004	0.794	0.016
200A_1000I	0.882	0.007	0.783	0.004	0.721	0.009	0.795	0.013
500A_100I	0.852	0.005	0.753	0.010	0.718	0.008	0.783	0.019
500A_200I	0.875	0.006	0.775	0.011	0.738	0.007	0.807	0.014
500A_500I	0.898	0.003	0.800	0.005	0.759	0.007	0.824	0.012
500A_1000I	0.909	0.004	0.815	0.005	0.759	0.007	0.828	0.012
1000A_100I	0.865	0.006	0.762	0.011	0.728	0.009	0.790	0.016
1000A_200I	0.885	0.004	0.790	0.010	0.747	0.005	0.815	0.014
1000A_500I	0.908	0.003	0.820	0.006	0.768	0.007	0.847	0.008
1000A_1000I	0.924	0.005	0.836	0.005	0.780	0.005	0.857	0.009

(F)

POLH	P/Z-Z		P/P–P		P/Z-P		P/P-Z	
# refs	AUC	SD	AUC	SD	AUC	SD	AUC	SD
100A_100I	0.931	0.005	0.830	0.012	0.824	0.006	0.903	0.011
100A_200I	0.937	0.004	0.837	0.011	0.829	0.008	0.906	0.010
100A_500I	0.942	0.007	0.848	0.005	0.828	0.010	0.911	0.006
100A_1000I	0.944	0.005	0.848	0.007	0.823	0.009	0.905	0.006
200A_100I	0.938	0.005	0.842	0.008	0.837	0.006	0.910	0.013
200A_200I	0.947	0.003	0.858	0.005	0.848	0.005	0.922	0.008
200A_500I	0.954	0.002	0.866	0.004	0.849	0.004	0.924	0.008
200A_1000I	0.955	0.002	0.869	0.005	0.843	0.004	0.924	0.006
500A_100I	0.943	0.004	0.848	0.005	0.839	0.007	0.913	0.008
500A_200I	0.954	0.002	0.865	0.004	0.853	0.003	0.927	0.007
500A_500I	0.962	0.001	0.881	0.003	0.862	0.004	0.938	0.003
500A_1000I	0.966	0.002	0.887	0.003	0.864	0.004	0.940	0.002
1000A_100I	0.947	0.005	0.853	0.006	0.842	0.008	0.922	0.007
1000A_200I	0.959	0.001	0.869	0.006	0.858	0.003	0.934	0.006
1000A_500I	0.968	0.002	0.891	0.003	0.870	0.004	0.947	0.004
1000A_1000I	0.973	0.002	0.900	0.002	0.874	0.003	0.950	0.004

[a] For each data set, average AUC values over 10 independent trials and standard deviations (SD) are reported for different numbers of active and inactive reference compounds (# refs) comprising all possible combinations of 100, 200, 500, and 1000 active (A) and inactive (I) compounds. For example, "100A_100I" refers to a training set consisting of 100 active and 100 inactive compounds. Training/test categories are abbreviated as

Table 2. continued

defined in the Materials and Methods section (P/P–P, P/P-Z, P/Z-P, and P/Z-Z). Virtual screening results are reported for all six data sets, and ECFP4 as the molecular representation: (A) EHMT2, (B) ALDH1A1, (C) BAZ2B, (D) Nrf2, (E) SMAD3, and (F) POLH.

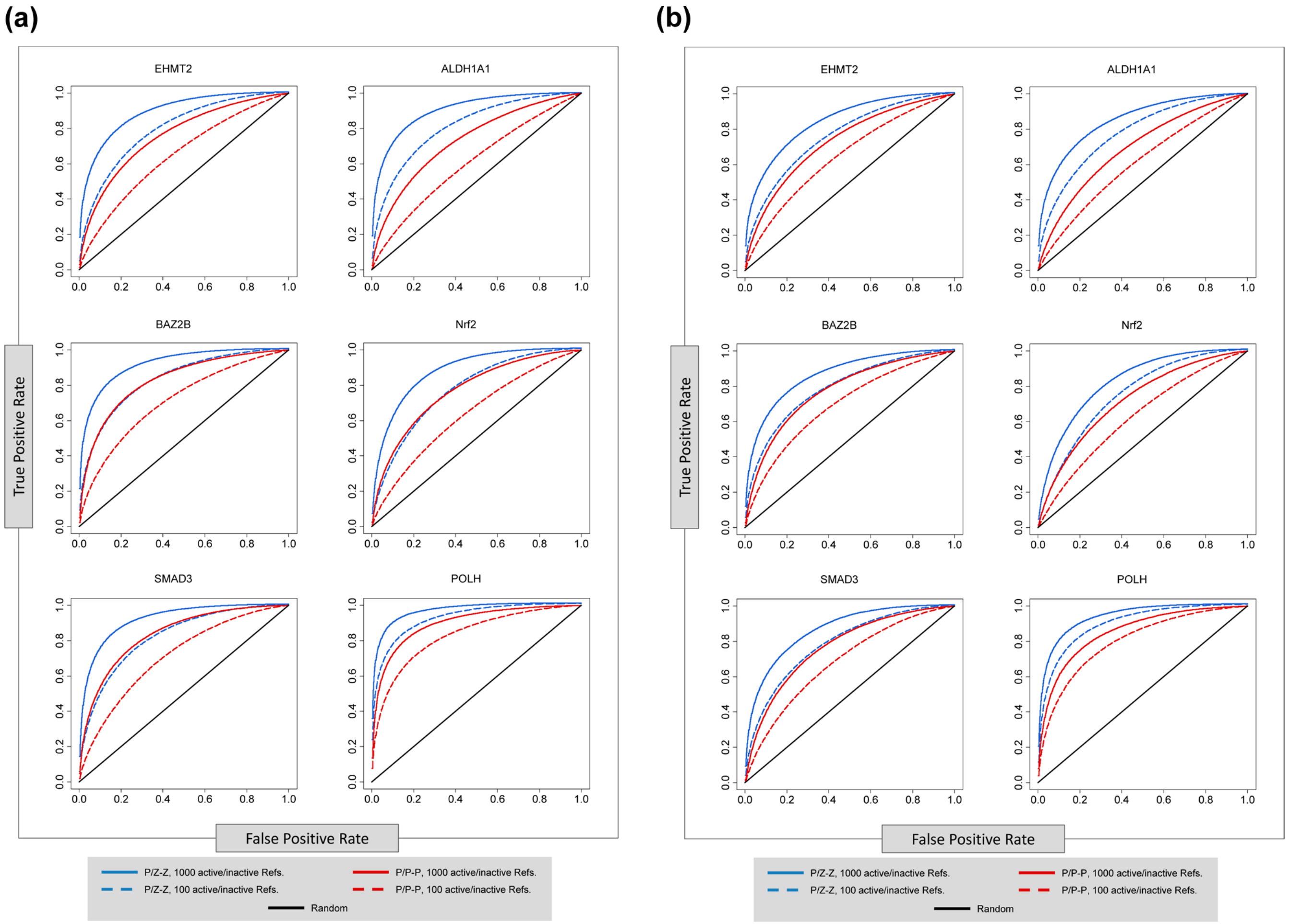


Figure 1. ROC curves. For all data sets, ROC average curves are shown for P/Z-Z (blue) and P/P–P (red) SVM calculations. In both cases, results are reported for 100 active and inactive reference compounds (dashed line) and for 1000 active and inactive training compounds (solid line). The black line represents random search performance. (a) ECFP4, (b) MACCS.

these two molecular representations, a total of 128 virtual screening constellations were obtained.

In each case, 10 different trials with randomly assembled positive and negative training and test sets were carried out. Virtual screening performance was measured using the receiver operating characteristic (ROC), and the area under the ROC curve (AUC)[27] averaged over all 10 trials.

All SVM calculations were carried out using $\text{SVM}^{\text{light}}$,[28] a freely available SVM implementation, and as calculation parameters, $\text{SVM}^{\text{light}}$ default settings were used.

■ RESULTS AND DISCUSSION

Study Goal and Design. The major focal point of our study is the question to what extent the choice of confirmed inactive molecules versus randomly selected database compounds as negative training examples might influence the outcome of SVM-based virtual screening calculations. Typically, the use of negative training data is not much discussed in SVM modeling and virtual screening applications. In benchmark calculations, it is common practice to use randomly selected database compounds, mostly from ZINC, as training compounds assumed to be inactive against a target of interest. This represents an approximation underlying model building, and it would scientifically be more rigorous to utilize confirmed inactive compounds for training. To address this issue, we have carried out systematic SVM calculations for different compound data sets obtained from PubChem Confirmatory Bioassays by applying well-defined training/test categories. These categories represented different combinations of PubChem and/or ZINC compounds for training and as the background database. This made it possible to compare SVM search performance for different training and test settings in detail.

Global Search Performance. Table 2 reports AUC values (and their standard deviations) for all SVM calculations using the ECFP4 fingerprint. The corresponding search results for MACCS structural keys are provided in Table S1 of the Supporting Information. In addition, in Figure 1A and B, ROC curves are shown for the smallest and largest training sets of P/Z-Z and P/P–P calculations for ECFP4 and MACCS, respectively.

Overall, the SVM-based virtual screening results yielded high search performance for the best categories, with AUC values of

~0.8−0.9 or even higher for all compound data sets. Highest search performance was generally observed for P/Z-Z calculations, i.e., when ZINC compounds were utilized as negative training examples and as the background database. Over all data sets, these calculations yielded AUC values that were on average ~0.1−0.2 higher than for P/P−P calculations, i.e., when confirmed inactive were used for training and search calculations were carried out in the PubChem screening database. These in part significant differences are reflected in Figure 1. For both fingerprint representations, equivalent trends were observed (Figure 1). AUC values were generally slightly higher for ECFP4 than for MACCS.

Training Set Composition. Increasing numbers of reference compounds generally led to increasing search performance, as illustrated in Figure 1. Average increases in AUC values ranged from ~0.04−0.11. In addition, with increasing numbers of reference compounds, standard deviations of the calculations decreased. However, no notable changes in AUC values were observed when training sets having the same size, but inverted composition were compared, e.g., 200 active and 500 inactive vs 500 active and 200 inactive compounds. Thus, overall increases in the number of reference compounds had a stronger influence on SVM performance than training set permutations.

Category-Dependent Differences in Search Performance. The different training/test categories we defined displayed systematic differences in search performance. As a general trend, search performance over all data sets decreased in the following order: P/Z-Z > P/P-Z > P/P-P > P/Z-P. Highest standard deviations were generally observed for P/P-Z calculations. The observed order indicated that confirmed active compounds were generally easier to identify on a ZINC background than in the screening database from which they originated, regardless of whether the SVM models were trained with PubChem or ZINC compounds as negative training instances. The most likely explanation for this finding is that active and inactive PubChem confirmatory assay compounds are often more similar to each other than active PubChem and random ZINC compounds, which are chemically very diverse. This explanation is also consistent with the observation that P/Z-Z calculations, which best exploited chemical differences between compounds gave the overall best search results. These calculations also produced AUC values that were on average ~0.10−0.21 higher than P/Z-P calculations. Furthermore, when PubChem compounds were used as negative training examples, the search performance was even slightly higher on a ZINC than a PubChem background, with average increases in AUC values of ~0.01−0.06. Thus, active compounds were easier to distinguish from ZINC compounds. These findings are also consistent with a major influence of the background database on the search results, irrespective of training conditions, with ZINC compounds yielding consistently best results.

Practical Implications. The results we obtained indicate that building predictive SVM models on the basis of actual screening libraries is more difficult than using combinations of known active and randomly chosen database compounds, although the use of confirmed inactive screening compounds as negative training examples is scientifically more accurate. By comparing the different screening setups explored, the background database was found to play a major role for the success of the virtual screening calculations. Simply put, confirmed active compounds were easier to distinguish from random ZINC selections than from the screening database from which they originated, indicating that many ZINC compounds might contain chemical characteristics that distinguish them from compounds selected for screening libraries. It should be noted that the P/Z-Z category represents a typical benchmark setting. In this case, models were derived from active and random training examples and used to screen a ZINC background database to with known actives were added. This setup produced consistently best results. For practical applications, a more realistic scenario would be to train SVM models on confirmed active and inactive compounds obtained from an initial experimental screen and apply these models to search a larger screening collection. This case exactly corresponds to our P/P−P category. However, under these conditions, search performance was consistently lower than for the P/Z-Z category. These findings indicate that SVM search performance under typical benchmark conditions is likely overestimated. In fact, active compounds utilized here were confirmed screening hits that were not chemically optimized. Hence, such screening hits differ from many optimized active compounds that are usually obtained from medicinal chemistry databases such as ChEMBL and used for benchmarking. Given their generally higher chemical complexity compared to screening hits, such active compounds are even easier to differentiate from random ZINC compounds, which can be expected to further increase search performance in benchmark calculations.

■ CONCLUSIONS

Herein, we have investigated the question to what extent the choice of negative training examples influences the outcome of SVM-based virtual screening. This question is usually only little considered in SVM modeling. For this purpose, we have compared a variety of SVM models that were generated on the basis of confirmed active and inactive compounds from different PubChem Confirmatory Bioassays with corresponding models generated from confirmed actives and randomly selected ZINC compounds. In these calculations, a clear influence of negative training examples on SVM search performance was detected. The results of search calculations using the same numbers of confirmed inactive training compounds and ZINC molecules assumed to be inactive systematically differed. In addition, we also observed that generally highest search performance was achieved when SVM models were screened on a ZINC background, regardless of whether the models were trained using inactive PubChem compounds or ZINC molecules. These findings revealed a strong influence of the background database on the virtual screening results. The best SVM models we obtained were derived from known active compounds and random ZINC collections and applied on a ZINC background, which corresponded to typical benchmark conditions.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Table S1 reports average AUC values and standard deviations for MACCS structural keys. This material is available free of charge via the Internet at http://pubs.acs.org.

■ 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) Geppert, H.; Vogt, M.; Bajorath, J. Current trends in ligand-based virtual screening: Molecular representations, data mining methods, new application areas, and performance evaluation. *J. Chem. Inf. Model.* **2010**, *50*, 205−216.

(2) Chen, B.; Harrison, R. F.; Papadatos, G.; Willett, P.; Wood, D. J.; Lewell, X. Q.; Greenidge, P.; Stiefl, N. Evaluation of machine-learning methods for ligand-based virtual screening. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 53−62.

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

(4) 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.

(5) 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.

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

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

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

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

(10) 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.

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

(12) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: A large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2012**, *40*, D1100−D1107.

(13) Liu, T.; Lin, Y.; Wen, X.; Jorisson, R. N.; Gilson, M. K. BindingDB: A Web-accessible database of experimentally determined protein−ligand binding affinities. *Nucleic Acids Res.* **2007**, *35*, D198−D201.

(14) Han, L. Y.; Ma, X. H.; Lin, H. H.; Jia, J.; Zhu, F.; Xue, Y.; Li, Z. R.; Cao, Z. W.; Ji, Z. L.; Chen, Y. Z. A support vector machines approach for virtual screening of active compounds of single and multiple mechanisms from large libraries at an improved hit-rate and enrichment factor. *J. Mol. Graph. Model.* **2008**, *26*, 1276−1286.

(15) Plewczynski, D.; Spieser, S. A. H.; Koch, U. Assessing different classification methods for virtual screening. *J. Chem. Inf. Model.* **2006**, *46*, 1098−1106.

(16) Gillet, V. J.; Willett, P.; Bradshaw, J. Identification of biological activity profiles using substructural analysis and genetic algorithms. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 165−179.

(17) Smusz, S.; Kurczab, R.; Bojarski, A. J. The influence of the inactives subset generation on the performance of machine learning methods. *J. Cheminf.* **2013**, *5*, 17 DOI: 10.1186/1758-2946-5-17.

(18) Irwin, J. J.; Sterling, T.; Mysinger, M. M.; Bolstad, E. S.; Coleman, R. G. ZINC: A free tool to discover chemistry for biology. *J. Chem. Inf. Model.* **2012**, *52*, 1757−1768.

(19) *Molecular Drug Data Report (MDDR)*; Accelrys, Inc., San Diego, CA. http://www.accelrys.com (accessed June 28, 2013).

(20) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking sets for molecular docking. *J. Med. Chem.* **2006**, *49*, 6789−6801.

(21) 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.

(22) 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.

(23) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.

(24) *MACCS Structural Keys*; Accelrys, San Diego, CA.

(25) *Molecular Operating Environment (MOE)*; Chemical Computing Group, Inc., Montreal, Quebec, Canada.

(26) Ralaivola, L.; Swamidass, S. J.; Saigo, H.; Baldi, P. Graph kernels for chemical informatics. *Neural Networks* **2005**, *18*, 1093−1110.

(27) Witten, I. H.; Frank, E. *Data Mining − Practical Machine Learning Tools and Techniques*, 2nd ed.; Morgan Kaufmann: San Francisco, 2005, pp 161−176.

(28) 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.