JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

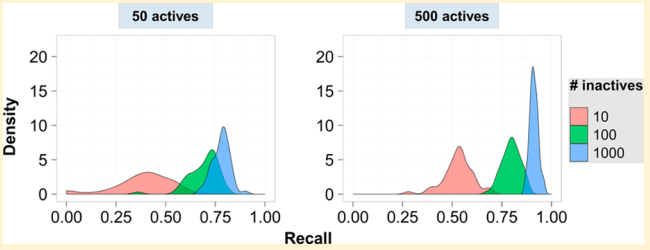
pubs.acs.org/jcim

Influence of Varying Training Set Composition and Size on Support Vector Machine-Based Prediction of Active Compounds

Raquel Rodríguez-Pérez, Martin Vogt, and Jürgen Bajorath*

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

ABSTRACT: Support vector machine (SVM) modeling is one of the most popular machine learning approaches in chemoinformatics and drug design. The influence of training set composition and size on predictions currently is an underinvestigated issue in SVM modeling. In this study, we have derived SVM classification and ranking models for a variety of compound activity classes under systematic variation of the number of positive and negative training examples. With increasing numbers of negative training compounds, SVM classification calculations became increasingly accurate and stable. However, this was only the case if a required threshold of positive training examples was also reached. In addition, consideration of class weights and optimization of cost factors substantially aided in balancing the calculations for increasing numbers of negative training examples. Taken together, the results of our analysis have practical implications for SVM learning and the prediction of active compounds. For all compound classes under study, top recall performance and independence of compound recall of training set composition was achieved when 250−500 active and 500−1000 randomly selected inactive training instances were used. However, as long as ∼50 known active compounds were available for training, increasing numbers of 500−1000 randomly selected negative training examples significantly improved model performance and gave very similar results for different training sets.



INTRODUCTION

The support vector machine (SVM) algorithm[1,2] is among the most widely used supervised machine learning methods in chemoinformatics and computer-aided drug discovery.[3−5] The popularity of SVM modeling primarily stems from generally high predictive performance in compound classification and virtual screening.[4] Although SVMs have been applied to investigate a variety of class label prediction and also regression tasks in chemoinformatics and drug discovery research,[4,5] so far only very few studies have addressed the issue of training set composition and size for SVM modeling[6] and other machine learning methods.[7,8] Especially the choice of negative training examples is often little considered in machine learning. Typically, to train models for compound classification, a subjectively chosen number of molecules are randomly selected from chemical databases to serve as negative training instances, without further analysis. Two previous studies have investigated the choice of negative training examples in greater detail.[6,7] For SVM modeling, the use of experimentally confirmed negative training compounds from screening assays and randomly chosen compounds from the ZINC database[9] was compared in the prediction of active compounds.[6] It was shown that the source of negative training instances affected the performance of SVM classification. Perhaps surprisingly, randomly selected ZINC compounds often resulted in better models than screening compounds that were confirmed to be inactive against a target for which active compounds were predicted.[6] No training set variations were carried out. In another study, negative training sets were assembled from different databases for compound classification using different machine learning approaches.[7] These calculations revealed a notable influence of negative training examples on the predictions and a preference for randomly selected ZINC compounds over compounds from other sources.[7] In this case, the size of negative training sets was varied when building models using different machine learning methods including SVMs with polynomial kernels. Training set size variations were found to influence compound predictions.[7] Performance relationships for varying numbers of negative and positive training examples were not investigated. In other studies, positive and negative training examples were balanced to improve the performance of machine learning models,[6,8] addressing the issue of data imbalance in machine learning.[10,11]

Herein, we report an analysis of the influence of training set composition and size on SVM classification and ranking by systematically varying the number of negative and positive training examples and determining how these variations affect the prediction of active compounds and stability of the calculations.

MATERIALS AND METHODS

SVM Classification. For SVM classification,[1] training compounds are defined by a feature vector $\mathbf{x} \in \mathcal{X}$ and a class label $\gamma \in \{-1, 1\}$ and projected into the reference space $\mathcal{X}$. SVMs solve a convex quadratic optimization problem to find a

Received: February 15, 2017
Published: April 4, 2017

DOI: 10.1021/acs.jcim.7b00088
J. Chem. Inf. Model. 2017, 57, 710−716

hyperplane $H = \{\mathbf{x}|\langle \mathbf{w}, \mathbf{x}\rangle + b = 0\}$ that separates the positive and negative class. The hyperplane H is defined by a normal vector $\mathbf{w}$ and a bias b and maximizes the margin between the two classes. To achieve model generalization, non-negative slack variables ξ_i are considered during training to penalize misclassification. In addition, the cost hyperparameter C controls the trade-off between margin maximization and permitted training errors, and its value can be optimized by cross-validation.[12]

Once the decision boundary is defined, test instances are projected into the feature space. New compounds of unknown class label are classified according to the side of the hyperplane on which they fall or, alternatively, ranked according to the value of $g(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x}\rangle$.[13] The latter strategy is equivalent to changing the bias of the hyperplane, sliding it from the most distant points on the positive side toward the negative side, and ranking compounds in the order they pass through the plane.

In the case of nonlinearly separable training data in a given reference space, the scalar product $<\cdot, \cdot>$ can be replaced by a kernel function $K(\cdot, \cdot)$, which is known as the *kernel trick*.[14] Using kernel functions, the scalar product of two feature vectors can be computed in a higher dimensional space $\mathcal{H}$ where the data may be linearly separable without the need to explicitly compute the mapping of $\mathcal{X}$ into $\mathcal{H}$. In SVM-based compound classification, the Tanimoto kernel is one of the most frequently used kernel functions for binary fingerprints.[15]

For imbalanced data sets, different class weights can be assigned to put relative weights on misclassification of positive and negative training instances and avoid orienting the hyperplane toward the minority class. Accordingly, C_+ and C_- balance the weight on slack variables for the positive and negative class, respectively.[16]

$$\frac{C_+}{C_-} = \frac{|\{i|y_i = -1\}|}{|\{i|y_i = +1\}|}$$

Compound Data Sets and Representation. Ten sets with at least 600 active compounds (positive instances) were obtained from ChEMBL version 22.[17] Only compounds with numerically specified equilibrium constants (K_i values) for single human proteins were selected, while omitting borderline active compounds ($pK_i < 5$) that might often represent artifacts. Table 1 reports the accession number, target name, number of compounds and mean pK_i values for these 10 compound data sets. As background set (pool of negative instances), 250 000 compounds were randomly selected from ZINC.[9] Random subsets of these compounds were used as negative training and test examples. For model building, all active and inactive compounds were represented as standard MACCS fingerprints[18] consisting of 166 bits monitoring the presence (bit set on) or absence (set off) of predefined structural fragments or patterns. Although we deliberately selected the simplistic and easy to rationalize MACCS fingerprint for our proof-of-concept investigation, control calculations were also carried out using the folded version of the extended connectivity fingerprint with bond diameter 4 (ECFP4).[19]

Table 1. Compound Data Sets[a]

accession no.	target name	number of compounds	mean pK_i
P00734	thrombin	839	6.67
P00918	carbonic anhydrase 2	2164	7.22
P21917	dopamine D4 receptor	804	7.11
P41146	nociceptin receptor	844	7.81
P00742	coagulation factor X	1476	7.77
P29275	adenosine receptor A2b	1187	7.12
P32245	melanocortin receptor 4	1260	7.00
Q9H3N8	histamine H4 receptor	875	6.97
Q99705	melanin-concentrating hormone receptor 1	1208	7.45
Q9Y5Y4	prostaglandin D2 receptor 2	833	7.53

[a]Ten compound data sets were selected from ChEMBL and used for SVM modeling. For each activity class, the ChEMBL accession no., target name, number of compounds, and mean pK_i value are reported.

Calculation Protocol.

(1) Each activity class was randomly divided into training and test (prediction) sets. Training set size was varied across values #I = {10, 50, 100, 500, 1000} for the negative (inactive) class and #A = {10, 50, 100, 250, 500} for the positive (active) class. Test sets always consisted of 10 000 inactive and 100 active compounds.

(2) Preprocessing of the fingerprints of the training and test data was carried out by removing zero-variance features and applying centering and unit variance scaling to all features on the basis of the training set for each trial.

(3) For each of the 25 training set combinations, SVM models were built using the linear and Tanimoto kernel with class weights C_+ and C_-. In addition, cost factors C controlling the influence of individual support vectors were optimized using values of 0.01, 0.1, 1, and 10. For cost factor optimization, 10-fold cross-validation was carried out with training data splits of 60% (model derivation) and 40% (testing, internal validation). Models with best cost factors were selected on the basis of largest area under the ROC curve (AUC).

(4) The optimized SVM model was used to rank test set compounds in the order of decreasing probability of activity based upon the signed distance from the hyperplane (positive to negative side). Model performance was assessed by determining the recall rate of active compounds within the top 1% of ranked test compounds. In addition, balanced accuracy (BA) was calculated, defined as

$$\mathrm{BA} = \frac{0.5\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} + \frac{0.5\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$$

(TP, true positives; TN, true negatives; FP, false positives; FN, false negatives).

(5) For each activity class and combination of a kernel function and training set size, the modeling process was carried out 50 times to obtain a distribution of recall rates.

(6) The results were compared using hypothesis testing. The nonparametric Kolmogorov–Smirnov test[20] was employed to account for differences between cumulative recall distributions and the Levene test[21] to compare the variance of these distributions. In addition, the Bonferroni correction[22] was introduced for multiple testing.

The calculation protocol was implemented in R,[23] and the *kernlab* package[24] was used for SVM modeling.

DOI: 10.1021/acs.jcim.7b00088
J. Chem. Inf. Model. 2017, 57, 710–716

RESULTS AND DISCUSSION

For different activity classes, SVM classification and ranking models were built under systematic variation of training set composition and size and active compounds were predicted. Specifically, the number of negative and positive training examples was varied in the ranges of 10−1000 and 10−500, respectively, and all possible combinations were explored. In addition, cost factors were optimized by cross-validation and class-specific weights were used to account for data imbalance in the training set.

Class Weights. Figure 1 compares balanced accuracy of the predictions in the presence or absence of class weights for two representative activity classes. Consideration of class-specific weights consistently improved the accuracy of the predictions for imbalanced training sets, except for three cases of large training sets with at least 250 actives and 500 inactives for which the performance was comparable. Hence, the explicit consideration of different class weights for positive and negative training instances produced more accurate classification models. Under these conditions, the derived hyperplane was not skewed toward the minority class, resulting in improved model generalization, especially in the presence of large training data imbalance. These effects were outweighed only for the largest and least imbalanced training sets. Given the demonstrated relevance of class weights for prediction accuracy, a factor that is not always considered in SVM modeling, results reported in the following included class weight settings.

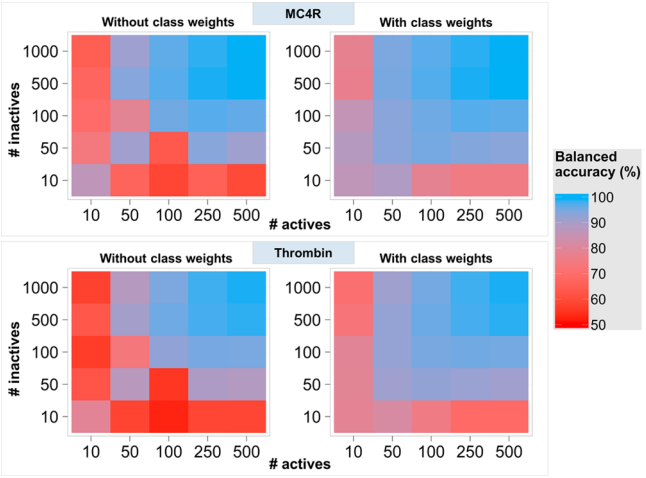


Figure 1. Effects of class weights on model performance. Heat map representations show balanced accuracy over 50 independent trials (using a two-color gradient) for training sets of varying composition and size: (top) melanocortin receptor 4 (MC4R) ligands, (bottom) thrombin inhibitors.

In addition, optimization of cost factors was carried out using cross validation. The best cost factors often varied depending on training set composition, but for well-performing training sets (i.e., those with large numbers of actives and inactives), there was an overall preference for C values of 0.01 for both the linear and Tanimoto kernels. For highly imbalanced data sets, larger cost factors were frequently selected, indicating that adjusting margin softness (stability) also contributed to model generalization. It is noteworthy that for different training set compositions and regardless of the cost factor chosen the hyperplanes generated by the SVMs were very frequently able to separate the training data without error and thus resulted in a hard margin classifier.

Kernels and Fingerprints. Figure 2 reports compound recall for alternative kernel functions under systematic variation of inactive and active training instances for two representative activity classes. Figure 3 shows corresponding density plots for recall rate distributions over multiple trials. First, we focus on relative kernel performance. The results in Figure 2 and 3 reveal generally higher recall performance for the Tanimoto than the linear kernel, frequently reaching a recall level of 0.9. However, even for the linear kernel, satisfactory recall was observed, often approaching a recall level of 0.75. Differences in recall performance between the linear and Tanimoto kernel were quantitatively assessed for all activity classes and statistically compared using the two-sided and paired Kolmogorov−Smirnov test. The results confirmed that the Tanimoto kernel generally performed significantly better than the linear kernel for training instances of #A = {100, 250, 500} and #I = {100, 500, 1000}. However, there was no significant difference in the cases of #A = {10} and #I = {50, 100, 500, 1000} where prediction accuracy was limited. Furthermore, as shown in Figure 3, SVM models derived using the Tanimoto kernel were generally more robust, i.e., corresponding recall rate distributions were sharper for the Tanimoto than for linear kernel. The presence of narrow distributions indicated that models derived from different training sets had comparable prediction accuracy for alternative test instances. As a control, SVM calculations were also repeated using the radial basis function (RBF) kernel,[25,26] another popular kernel function, with a sigma setting, corresponding to the inverse kernel width, of 0.01.[26] The results obtained using the RBF kernel were, on average, nearly indistinguishable from those obtained using the Tanimoto kernel discussed in the following. As an additional control, the calculations were also carried out using ECFP4 instead of MACCS to compare the trends observed for training set variation. With both fingerprints, the same trends were observed (with the typical slightly better recall performance of ECFP4 relative to MACCS).

Training Sets of Varying Composition and Size. The results in Figures 2, 3, and 4 revealed two key findings; (i) recall performance and model generalization consistently improved with increasing size of training sets and (ii) the ratio of active vs inactive training examples significantly influenced prediction accuracy. The increases in recall performance observed in Figure 2 were detected for all activity classes. When the number of active training instances was kept constant, recall rates increased with increasing numbers of inactive instances, except in the case of 10 actives, where prediction accuracy was generally low even over the range of 100−1000 negative instances. Thus, a minimum number of active training compounds was required for training sets of increasing size. Similar observations were made when the number of inactive training compounds was kept constant and the number of active examples was increased. Ten negative examples were consistently insufficient for building effective models and 50 negative training instances were often insufficient (Figure 2). However, in the presence of at least 100 negative training instances, high prediction accuracy was consistently achieved when the number of active examples was increased (Figure 3).

For all compound classes, incremental increase in the number of negative (positive) training instances led to systematic performance enhancements when at least 50 positive (100 negative) training compounds were used, as confirmed by

DOI: 10.1021/acs.jcim.7b00088
J. Chem. Inf. Model. 2017, 57, 710−716

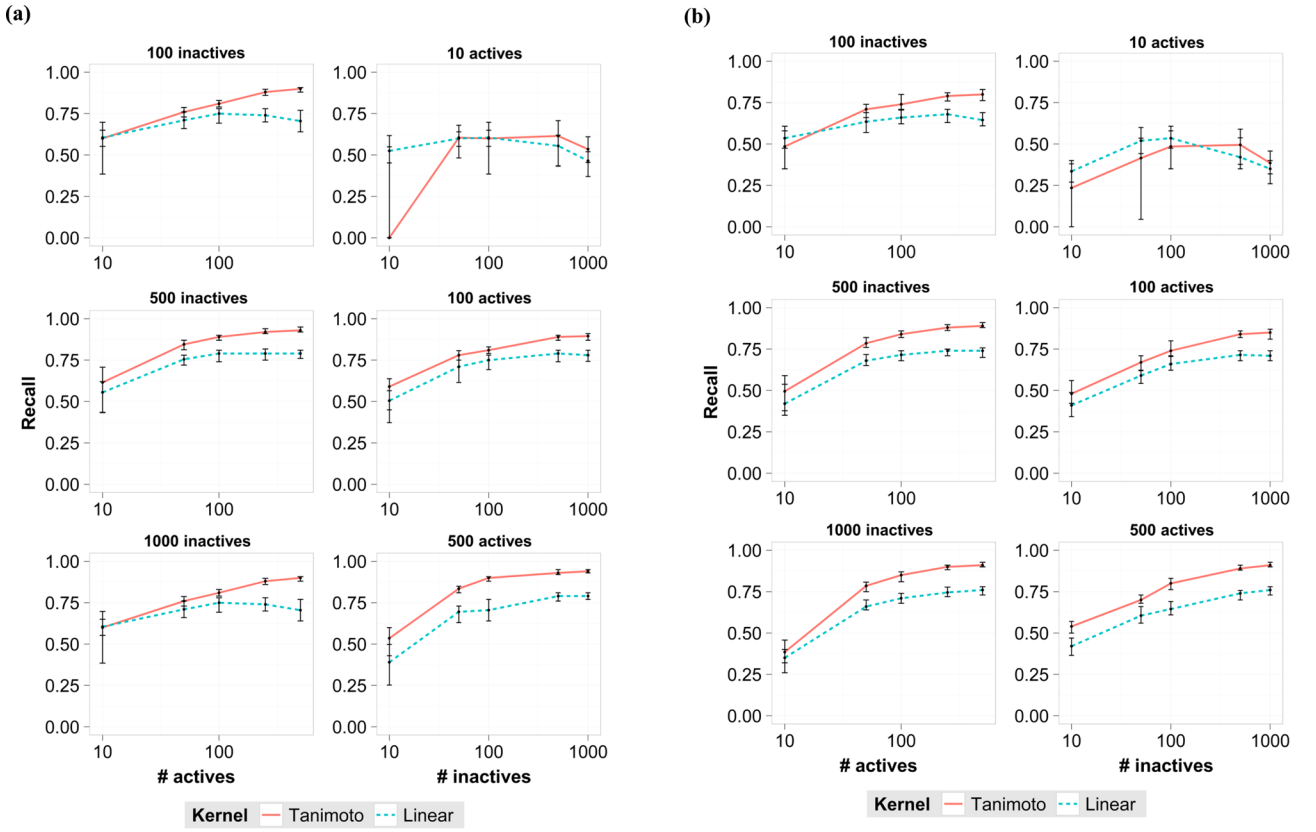


Figure 2. Recall performance. The median value and interquartile range of the recall rate of active compounds among the top 1% of the ranking is reported for 50 trials with the linear (blue dashed line) or Tanimoto (red solid line) kernel. Results monitor the evolution of recall for a constant number of inactives (or actives) and increasing number of actives (or inactives) in the training set: (a) melanocortin receptor 4 ligands, (b) thrombin inhibitors.

the one-sided Kolmogorov−Smirnov test. While overall highest prediction accuracy was achieved for training sets consisting of 500 active and 1000 inactive examples, similar accuracy was already observed for 100 active and 500 inactive training compounds. Furthermore, recall generally began to reach a plateau when at least 100 active and 500 inactive training instances were used (Figure 2). However, with further increasing training set size, recall rate distributions became narrower, as illustrated in Figure 3 and 4, which was indicative of models with consistent prediction accuracy despite training set variations, as mentioned above.

Table 2 compares the recall performance over all activity classes for one of the worst and the best performing training set compositions of 10 actives/100 inactives and 500 actives/1000 inactives, respectively. In the bad case scenario, recall rates of compounds were—with one exception—lower than 50% with large standard deviations and balanced accuracy was around the 80% level. By contrast, for the best performing large training sets, recall rates were consistently high, with a mean of 87%, and balanced accuracy was approaching 100% with very low standard deviations (Table 2). Interestingly, training set imbalance only limited the accuracy of predictions in the case of small but not large training sets, as illustrated in Figure 4, an effect that can be ascribed to the use of class weights for SVM models, as detailed above. For example, while an inactive vs active ratio of 10:1 produced inaccurate predictions for training sets comprising 100 inactive and 10 active training examples, prediction accuracy was high when 1000 inactive and 100 active training compounds were used. Similar observations were made for other compound ratios.

Variance. Taken together, the results in Figure 3 and 4 clearly indicate that the predictions became stable with increasing size of training sets, another key finding. Figure 5 reports the variance of recall rates over independent predictions using training sets of increasing size and provides confirmatory evidence. Furthermore, Levene tests for all activity classes confirmed that the variance of recall distributions significantly differed in 38 of 40 cases (resulting from 10 compound classes and four training set conditions) when training sets with at least 50 active and 10 or 1000 inactive examples were used. By contrast, no statistically significant differences in variance of recall rate distributions were detected when the SVM models were trained with 100 or 1000 inactive examples, regardless of the number of actives.

■ CONCLUSIONS

Herein, we have systematically analyzed the influence of training set composition and size on the prediction accuracy of SVM classification models. Different from earlier studies, our calculations have stressed the importance of considering class weights and optimizing cost factors when imbalanced training sets are used. Furthermore, the ratio of active vs inactive training examples substantially affected the ability of SVM models to correctly predict active compounds. However, recall rates and balanced accuracy consistently improved for training sets of increasing size for all compound classes under study. Increasing size of training sets also compensated for inherent

DOI: 10.1021/acs.jcim.7b00088
J. Chem. Inf. Model. 2017, 57, 710−716

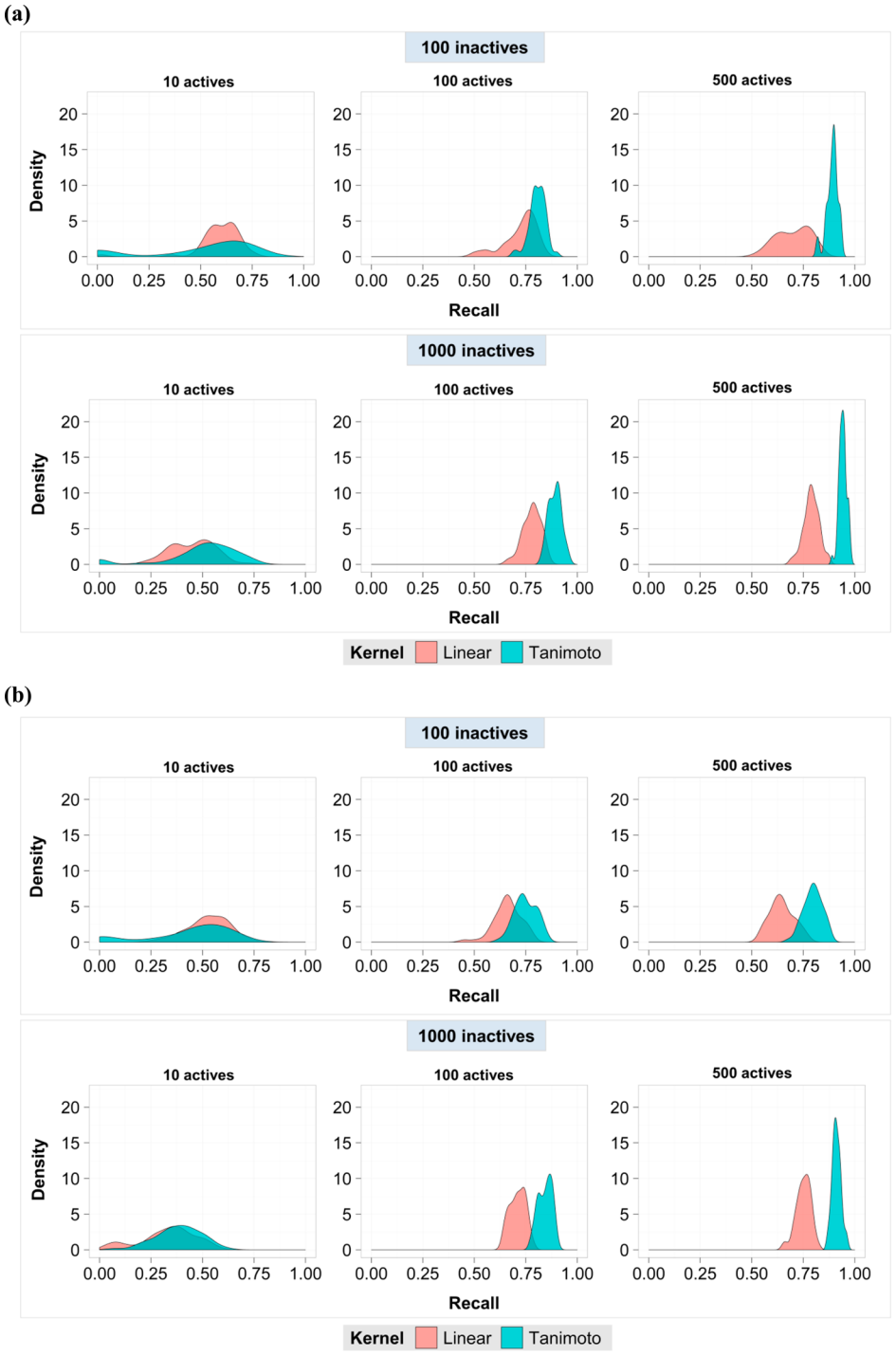


Figure 3. Density estimates. The distribution of recall rates over 50 trials is given for 100 (top) and 1000 (bottom) inactive and increasing numbers of active training compounds: (a) melanocortin receptor 4 ligands, (b) thrombin inhibitors.

data imbalance. Moreover, large training sets led to robust predictions and the accuracy was essentially constant when different training sets of the same size were used. Taken together, our findings have implications for practical applications of SVM classifiers. The following conclusions can be drawn. Best performing SVM models were obtained when 250−500 active and 500−1000 randomly selected inactive training instances were used. Moreover, as long as ∼50 known active compounds are available for training, increasing numbers of 500−1000 randomly selected negative training examples improve and stabilize model performance when class weights are taken into consideration, which provides a clear guideline for virtual compound screening.

Finally, we note that large numbers of active compounds may not always be available for training. However, since SVM classification and ranking models do not take compound potency as a parameter into account, in contrast to support vector regression, large numbers of hits often obtained from confirmatory screening assays might be readily used for SVM model building.

DOI: 10.1021/acs.jcim.7b00088
J. Chem. Inf. Model. 2017, 57, 710−716

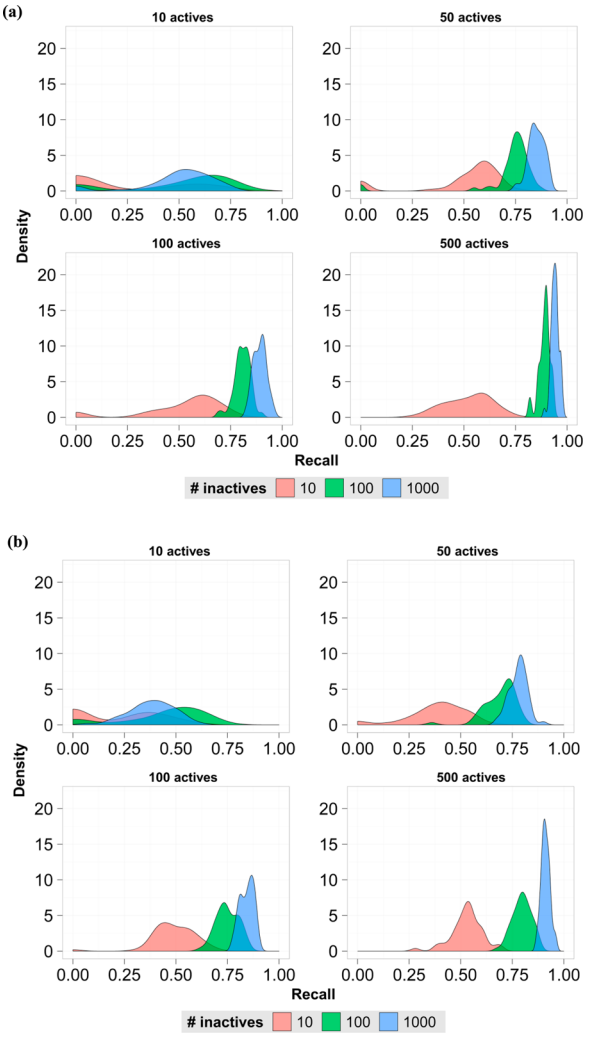


Figure 4. Influence of training set composition and size on recall rates. Density estimates obtained from the distribution of recall rates over 50 trials are presented for training sets of varying size and composition. For a constant number of 10−500 active training compounds, recall distributions are shown for 10 (pink), 100 (green), and 1000 (blue) inactive training compounds: (a) melanocortin receptor 4 ligands, (b) thrombin inhibitors.

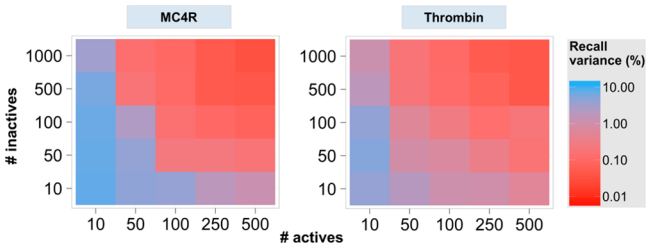


Figure 5. Influence of training set composition and size on recall variance. Heat map representations show variance of recall rates over 50 independent trials (using a two-color gradient) for training sets of varying composition and size: (left) melanocortin receptor 4 (MC4R) ligands, (right) thrombin inhibitors.

AUTHOR INFORMATION

Corresponding Author

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

ORCID

Jürgen Bajorath: 0000-0002-0557-5714

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The project leading to this report has received funding (for R.R.P.) from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 676434, "Big Data in Chemistry" ("BIGCHEM", http://bigchem.eu). The article reflects only the authors' view and neither the European Commission nor the Research Executive Agency (REA) are responsible for any use that may be made of the information it contains.

ABBREVIATIONS

AUC, area under receiver operating characteristic curve; BA, balanced accuracy; ECFP, extended connectivity fingerprint; MC4R, melanocortin receptor 4; RBF, radial basis function; SVM, support vector machine

Table 2. Classification Performance[a]

	10 actives and 100 inactives				500 actives and 1000 inactives			
accession no.	recall μ	recall σ	BA (%) μ	BA (%) σ	recall μ	recall σ	BA (%) μ	BA (%) σ
P00734	0.433	0.211	79.3	5.1	0.911	0.021	98.8	0.6
P00918	0.388	0.219	87.2	3.7	0.770	0.036	97.0	0.9
P21917	0.288	0.164	80.9	5.9	0.744	0.045	96.9	1.1
P41146	0.455	0.163	80.9	6.3	0.924	0.018	99.4	0.3
P00742	0.236	0.138	72.4	5.7	0.872	0.027	98.5	0.6
P29275	0.407	0.226	81.5	4.5	0.820	0.030	97.0	1.1
P32245	0.486	0.276	85.6	4.5	0.942	0.018	99.0	0.5
Q9H3N8	0.440	0.233	84.3	4.5	0.888	0.030	98.4	0.7
Q99705	0.349	0.171	78.7	6.9	0.860	0.046	98.2	0.7
Q9Y5Y4	0.562	0.206	83.7	4.8	0.965	0.013	99.3	0.6
global performance	0.405	0.200	81.4	5.2	0.870	0.028	98.2	0.7

[a]Reported are the mean (μ) and standard deviation (σ) of recall of active compounds and balanced accuracy after 50 independent trials for differently composed training sets: "10 active and 100 inactive compounds" (low performance) and "500 active and 1000 inactive compounds" (high performance). Results are shown for 10 compound classes, referred by accession no., according to Table 1. In addition, global performance over all classes is reported.

DOI: 10.1021/acs.jcim.7b00088
J. Chem. Inf. Model. 2017, 57, 710−716

REFERENCES

(1) Vapnik, V. N. *The Nature of Statistical Learning Theory*, 2nd ed.; Springer: New York, 2000.
(2) Burges, C. J. C. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Min. Knowl. Discovery* **1998**, *2*, 121−167.
(3) 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.
(4) 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.
(5) Heikamp, K.; Bajorath, J. Support Vector Machines for Drug Discovery. *Expert Opin. Drug Discovery* **2014**, *9*, 93−104.
(6) Heikamp, K.; Bajorath, J. Comparison of Inactive and Randomly Selected Compounds as Negative Training Examples in Support Vector Machine-Based Virtual Screening. *J. Chem. Inf. Model.* **2013**, *53*, 1595−1601.
(7) 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.
(8) Kurczab, R.; Smusz, S.; Bojarski, A. J. The Influence of Negative Training Set Size on Machine Learning-Based Virtual Screening. *J. Cheminf.* **2014**, *6*, 32.
(9) 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.
(10) Japkowicz, N. Learning from Imbalanced Data Sets: A Comparison of Various Strategies. *AAAI Workshop on Learning from Imbalanced Data Sets*, 2000; Vol. *68*, pp 10−15.
(11) Japkowicz, N.; Stephen, S. The Class Imbalance Problem: A Systematic Study. *Intell. Data Anal.* **2002**, *6*, 429−449.
(12) Stone, M. Cross-Validatory Choice and Assessment of Statistical Predictions. *J. R. Stat. Soc. Series B Stat. Methodol.* **1974**, *36*, 111−147.
(13) 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.
(14) 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, Pennsylvania; ACM: New York, 1992; pp 144−152.
(15) Ralaivola, L.; Swamidass, S. J.; Saigo, H.; Baldi, P. Graph Kernels for Chemical Informatics. *Neural Netw.* **2005**, *18*, 1093−1110.
(16) Morik, K.; Brockhausen, P.; Joachims, T. Combining Statistical Learning with a Knowledge-based Approach−A Case Study in Intensive Care Monitoring. In *Proceedings of the 16th International Conference on Machine Learning (ICML-99)*; Morgan Kaufmann: Burlington, MA, 1999.
(17) Bento, A. P.; Gaulton, A.; Hersey, A.; Bellis, L. J.; Chambers, J.; Davies, M.; Krüger, F. A.; Light, Y.; Mak, L.; McGlinchey, S.; et al. The ChEMBL Bioactivity Database: An Update. *Nucleic Acids Res.* **2014**, *42*, 1083−1090.
(18) *MACCS Structural keys*; Accelrys: San Diego, CA, USA, 2006.
(19) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.
(20) Dodge, Y. *The Concise Encyclopedia of Statistics*; Springer: New York, 2008.
(21) Brown, M. B.; Forsythe, A. B. Robust Tests for the Equality of Variances. *J. Am. Stat. Assoc.* **1974**, *69*, 364−367.
(22) Shaffer, J. P. Multiple Hypothesis Testing. *Annu. Rev. Psychol.* **1995**, *46*, 561−584.
(23) *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria.
(24) Karatzoglou, A.; Smola, A.; Hornik, K.; Zeileis, A. kernlab: An S4 Statistical Package for Kernel Methods in R. *J. Stat. Softw.* **2004**, *11*, 1−20.
(25) Amari, S. I.; Wu, S. Improving Support Vector Machine Classifiers by Modifying Kernel Functions. *Neural Netw.* **1999**, *12*, 783−789.
(26) Alvarsson, J.; Eklund, M.; Engkvist, O.; Spjuth, O.; Carlsson, L.; Wikberg, J. E.; Noeske, T. Ligand-Based Target Prediction with Signature Fingerprints. *J. Chem. Inf. Model.* **2014**, *54*, 2647−2653.

NOTE ADDED AFTER ASAP PUBLICATION

This article was published ASAP on April 10, 2017, with an error in the formula on page B, left column, second paragraph. The corrected version was published ASAP on April 11, 2017.

DOI: 10.1021/acs.jcim.7b00088
J. Chem. Inf. Model. 2017, 57, 710−716