Journal of
Chemical Information
and Modeling

Article

pubs.acs.org/jcim

Benchmarking Study of Parameter Variation When Using Signature Fingerprints Together with Support Vector Machines

Jonathan Alvarsson,*[†] Martin Eklund,[†,‡] Claes Andersson,[§] Lars Carlsson,[∥] Ola Spjuth,[†,⊥]
and Jarl E. S. Wikberg[†]

[†]Department of Pharmaceutical Biosciences, Uppsala University, SE-751 24 Uppsala, Sweden

[‡]Department of Surgery, University of California at San Francisco (UCSF), San Francisco, California 94115, United States

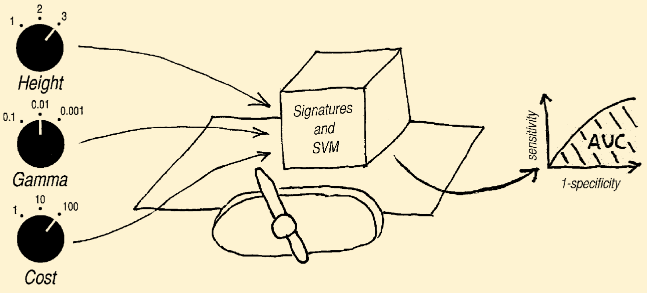
[§]Department of Medical Sciences, Uppsala University, SE-751 85 Uppsala, Sweden

[∥]Computational ADME and Safety, DSM, AstraZeneca R&D, SE-431 83 Mölndal, Sweden

[⊥]Science for Life Laboratory, Uppsala University, SE-751 24 Uppsala, Sweden

S Supporting Information

ABSTRACT: QSAR modeling using molecular signatures and support vector machines with a radial basis function is increasingly used for virtual screening in the drug discovery field. This method has three free parameters: C, γ, and signature height. C is a penalty parameter that limits overfitting, γ controls the width of the radial basis function kernel, and the signature height determines how much of the molecule is described by each atom signature. Determination of optimal values for these parameters is time-consuming. Good default values could therefore save considerable computational cost. The goal of this project was to investigate whether such default values could be found by using seven public QSAR data sets spanning a wide range of end points and using both a bit version and a count version of the molecular signatures. On the basis of the experiments performed, we recommend a parameter set of heights 0 to 2 for the count version of the signature fingerprints and heights 0 to 3 for the bit version. These are in combination with a support vector machine using C in the range of 1 to 100 and γ in the range of 0.001 to 0.1. When data sets are small or longer run times are not a problem, then there is reason to consider the addition of height 3 to the count fingerprint and a wider grid search. However, marked improvements should not be expected.



INTRODUCTION

In quantitative structure−activity relationships (QSARs), the activities of known molecules are modeled by describing the molecules numerically and correlating the numerical descriptions to the activities.[1] The QSAR models are subsequently used to predict the activities of new molecules whose activities are unknown. One modeling approach that is increasingly used in QSAR is the combination of molecular signatures[2−4] with support vector machines (SVMs).[5,6] Molecular signatures are topological descriptors that describe, for each atom in a molecule, the neighboring atoms in a canonical fashion. They have been used for natural-product-likeness scoring,[7] together with SVM in QSAR,[8−11] in molecular design for solvent selection,[12] and in modeling cytochrome P450 inhibition,[13,14] and they have been shown to produce robust models for drug design.[15] Norinder et al.[16] used molecular signatures together with both SVM and random forest when introducing conformal prediction in predictive modeling. We recently studied the molecular signature fingerprint and concluded that it performed on par with other well-established molecular fingerprints when it was compared to a set of other fingerprints by doing ligand-based target prediction based on k nearest neighbors by Tanimoto distance.[17]

The combination of signatures and SVM with the radial basis function (RBF) kernel constitutes a method where three parameters need to be determined before a model can be built and used to predict future observations. The molecular signature descriptor has one parameter, the height, which is the distance from a central atom. Rostkowski et al.[13] reported that for their data set the most successful combination of signature heights was 0 to 3, and Lapins et al.[14] reported that the combination of heights 1 to 3 gave the best SVM model for their data. However, neither of them reported which heights were investigated.

SVM with the RBF kernel has two free parameters that need to be determined, namely, C and γ. C is a penalty parameter that limits overfitting, and γ controls the width of the RBF kernel. One approach for determining the best values for these parameters is to perform a grid search by cross-validation on the training set.[18] In the grid search, a large number of SVM

Received: June 11, 2014

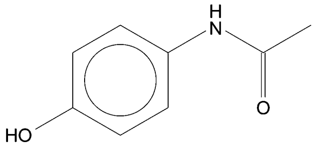
models with different values of C and γ are built on a training set, and the values leading to the best model are then used in building the final model. When new data become available over time, the models should be retrained.[19] If general default values for C and γ cannot be found, then the complete grid search should be redone. Grid searching is a computationally expensive method, and any hints as to which part of the grid should be searched could save considerable time both when building new models and when retraining old ones for new data. An approximation of these values could thus provide a good starting point for study-specific optimizations.

The aims of this project were first to benchmark what signature height is a good choice and second to study the effect on prediction performance of varying the SVM parameters C and γ by studying the average performance on a wide range of QSAR data sets.

■ METHODS

Support Vector Machines. SVM is a popular approach for solving classification and regression problems. For classification, the algorithm maps the problem into a high-dimensional space and chooses a hyperplane that separates the two classes.[5] The RBF kernel, used in this study, performs this mapping in a nonlinear fashion and is a commonly used kernel that has been suggested as a good starting point for SVM modeling.[20]

Signature Fingerprint. We used the signature fingerprints implemented in the open-source cheminformatics library Chemistry Development Kit (CDK).[21,22] CDK has one bit and one count signature fingerprint based on the signature descriptor.[17] An atom signature is a canonical string representation of the environment around an atom. The height variable determines how much of the atom environment is described (i.e., the number of atoms away from the center atom) (Figure 1). The atom signatures for all of the atoms in the molecule are hashed down to 32-bit integers, and the bit fingerprint simply tracks whether or not a certain signature hash value exists in the molecule. The count version keeps track of how many times the hash value exists in the molecule. Multiple previous studies of signature usage report tested height variations using a cumulative approach, i.e., first height 0, then height 0 and 1, followed by height 0 to 2, and so on.[9,13,15] This is a pragmatic approach since each additional height adds a great number of signature hash values. We used this cumulative approach and added one height at a time until no significant difference in performance could be detected. However, when the signature that contributes the most to a prediction is calculated and visualized as a substructure (as is done in Bioclipse DS[8,9]), it is easier to interpret the results without height 0 included, as this would correspond to highlighting individual atom types. Hence, we also looked at removing the lowest heights one at a time to see whether or not they made a relevant contribution to the performance.



Height	Signature for the nitrogen atom
0	[N]
1	[N]([C][C])
2	[N]([C](p[C]p[C])[C]([C]=[O]))

Figure 1. Example of atom signatures of height 0 to 2 for the nitrogen in acetaminophen (paracetamol). At height 0, the signature specifies only the atom type. As the height increases, more information on the atom environment is included. Reprinted from ref 17. Copyright 2014 American Chemical Society.

Data. We studied seven public data sets that have previously been used[23] for benchmarking in ligand-based modeling: COX2, DHFR, CPDB, EPAFHM, FDA, cas_N6512, and screen_U251. The COX2 data set[24,25] contains known cyclooxygenase-2 inhibitors, which are associated with treatment of inflammation, pain, and fever. The DHFR data set[24] contains inhibitors for the enzyme dihydrofolate reductase, which when inhibited causes DNA synthesis disruption.[26] The CPDB data set[27−29] contains carcinogenicity data. The EPAFHM data set[30] contains acute toxicology data measured on the Fathead Minnow fish. The FDA data set[31] concentrates on maximum recommended daily therapeutic doses. The cas_N6512 data set[32] contains mutagenicity data taken by the Ames assay. The screen_U251 data set[33] contains data from human tumor cell line screening. Thus, the data sets used span a wide range of use cases. Table 1 lists the data set sizes.

Table 1. Statistics for the Datasets: Number of Substances Classified as Positive and Total Number of Substances in the Dataset

	positive substances	substances
COX2	161 (50%)	322
DHFR	198 (50%)	397
CPDB	567 (47%)	1198
EPAFHM	289 (50%)	577
FDA	608 (50%)	1216
cas_N6512	3503 (54%)	6512
screen_U251	2033 (54%)	3743

Study Design. The experiment can be regarded as a five-factor factorial design with the following factors: C, γ, data set, height, and fingerprint type (bit/count). The data sets made up a series of *case-control studies*. The term "case-control study" originates from the medical field. A set of entities having a property is found, and a set of suitable controls is created. This means that the prevalences in the data sets are artificially determined and thus that measures based on the prevalence should not be used. However, from a case-control study it is possible to estimate the sensitivity and specificity, and from those values it is possible to create a receiver operating characteristic (ROC) curve. Thus, we calculated the area under the ROC curve (AUC)[34] for each parameter set. In order to get representative values for the different heights, a large number of C and γ values were tested. We built SVM models for each combination of C, γ, and height. The SVM models were evaluated using 10-fold cross-validation. The mean AUC value from all of the cross-validation folds was stored.

Table 2 shows the tested values for the factors. Let $\mathcal{C}$ be the set of the 12 values for C, $\mathcal{G}$ be the set of the 12 values for γ, $\mathcal{D}$ be the set of the seven data sets, $\mathcal{H}$ be the set of the six different height combinations, and $\mathcal{T}$ be the set of the two fingerprint types tested (i.e., bit and count). Our data can then be regarded as being of the form AUC_{cgdht}, where c, g, d, h, and t belong to the sets $\mathcal{C}$, $\mathcal{G}$, $\mathcal{D}$, $\mathcal{H}$, and $\mathcal{T}$, respectively.

Box plots showing the maximum AUC values upon variation of C and γ for each height were created separately for the bit

Table 2. Factor Levels Tested in the Benchmarking Experiments

C ($\mathcal{C}$)	γ ($\mathcal{G}$)	data set ($\mathcal{D}$)	height ($\mathcal{H}$)	fingerprint type ($\mathcal{T}$)
0.1	10	COX2	0	bit
1	1	DHFR	0−1	count
10	0.1	CPDB	0−2	
100	0.01	EPAFHM	0−3	
1×10^3	1×10^{-3}	FDA	0−4	
1×10^4	1×10^{-4}	cas_N6512	0−5	
1×10^5	1×10^{-5}	screen_U251		
1×10^6	1×10^{-6}			
1×10^7	1×10^{-7}			
1×10^8	1×10^{-8}			
1×10^9	1×10^{-9}			
1×10^{10}	1×10^{-10}			

and count fingerprints. The data in the box plots can be described by

$$\mathrm{AUC}_{dht}^{\max} = \max_{\forall(c,g)}\{\mathrm{AUC}_{cgdht}\} \tag{1}$$

Heat maps showing the AUC values for different values of C and γ for each data set and height were also generated. These heat maps were then summarized into two different types of aggregate heat maps. The first was obtained by calculating the mean AUC values from all of the data sets and heights for each combination of C and γ. Thus, each cell in the mean AUC heat maps can be described by

$$\mathrm{AUC}_{cgt}^{\mathrm{mean}} = \frac{\sum_{\forall(d,h)} \mathrm{AUC}_{cgdht}}{|\mathcal{D}|\cdot|\mathcal{H}|} \tag{2}$$

The second type of heat map was constructed by calculating how many times each combination of C and γ performed the best. Thus, each cell in the maximum AUC heat maps can be described by

$$N_{cgt}^{\max} = \sum_{\forall(d,h)} [\mathrm{AUC}_{cgdht} = \max_{c,g}\{\mathrm{AUC}_{cgdht}\}] \tag{3}$$

where

$$[P] = \begin{cases} 1 & \text{if } P \text{ is true} \\ 0 & \text{otherwise} \end{cases}$$

in which P is a statement that can be true or false.

A series of statistical tests were performed in order to answer two specific questions:

1. At which cumulatively added height does adding another height not yield better models?
2. On the basis of AUC values, can we recommend the use of either the count or the bit version of the fingerprint?

We computed the means and 95% confidence intervals of differences in AUC when signature heights were systematically added in order to assess the effect on the predictive performance. In order to say at what height the new result was not relevantly better than an older result, a nonsuperiority test was performed (cf. noninferiority testing[35]). A nonsuperiority test was chosen since a standard superiority test can test only for significant differences and not whether there is a relevant difference in effect (lack of significance does not mean that there is no effect but merely means that we lack the statistical power to detect any difference). In order to perform a nonsuperiority test, a threshold for what is to be considered a relevant difference has to be chosen. The choice of such a threshold depends on the situation and should preferably come from domain knowledge. We here somewhat arbitrarily decided that a difference of at least 0.01 in AUC on average was a relevant difference.

Finally, in order to study the effects of removing lower heights, another series of Wilcox tests were constructed using C and γ from a 3 × 3 grid search around the best area found in the heat maps and removing the lowest heights one at a time. The same threshold level of 0.01 was used in those tests.

The numbers of unique signatures used in modeling the different data sets at the different heights are given in Table 3.

Software. Signature fingerprints were calculated using CDK. SVM models were created in the programming language R[36] using the package e1071.[37] AUC values were calculated using the R package pROC.[38] The R package ggplot2[39] was used to create box plots. Heat maps were created with the R package gplots.[40]

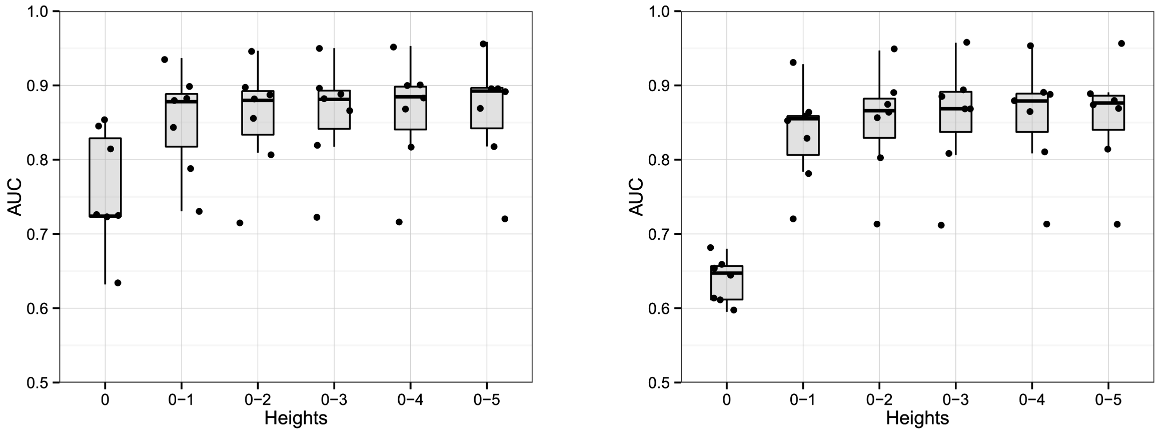
RESULTS

Box plots showing the 84 maximum AUC values are presented in Figure 2. It can be seen that the models do not perform very well with only signature height 0, especially not for the bit version. However, already with the addition of height 1 there is a leap in performance.

Figure 3 summarizes the heat maps by showing the mean AUC values as well as the number of times each combination of C and γ turned out to be best, as described by eqs 2 and 3, respectively. (All 84 generated heat maps are included in the Supporting Information.) From Figure 3 it can be seen that for both the count and bit versions of the fingerprint, the combination $C = 10$ and $\gamma = 0.01$ performed best with respect to both average AUC and the number of times the combination was the best. The second “hottest” cells in the two bottom heat

Table 3. Numbers of Unique Signatures Used To Represent the Different Datasets at the Tested Heights

heights	COX2	DHFR	CPD	EPAFHM	FDA	cas_N6512	screen_U251
0	8	8	25	10	10	13	34
0−1	99	84	359	162	257	461	560
0−2	646	551	3303	1201	3771	7742	7933
0−3	1909	1929	10522	3557	14132	34512	33312
0−4	3957	4483	20223	6313	29547	79637	73235
0−5	6688	8114	30432	8687	47610	134256	121201
1−3	1901	1921	10505	3547	14122	34499	33278
2−3	1810	1845	10237	3410	13878	34131	32767
3−3	1263	1378	7568	2509	10395	27536	25460



(a) AUCs for different heights of the bit fingerprint

(b) AUCs for different heights of the count fingerprint

Figure 2. Box plots showing the 84 maximum AUC values out of the mean AUCs from the cross-validation for the different heights as described by eq 1. The dots are somewhat spread out using a random jitter in order to reduce overlap. When only height 0 is used, neither the bit version (a) nor the count version (b) performs well. Each height addition increases the information available, but a saturation situation seems to occur where the information addition in each added height no longer leads to a better result.

maps in Figure 3 reside next to the cell for $C = 10$ and $\gamma = 0.01$, indicating a maximum in that area of the heat map.

Out of 6048 tested combinations of C, γ, and height, 76 (1.3%) were missing because those computations did not finish within a reasonable time. Most of the missing values were for height 0 to 1, so we regard the effect of these missing values as irrelevant for the end result since it is not at the lowest heights that we expect to find our region of greatest interest (see Figure 2). However, for the cas_N6512 data set one missing value occurred for height 2. The same combination of C and γ was also missing for height 1 for that data set, indicating that this combination was especially time-consuming. However, the missing combination is far from the sweet spot that repeatedly was found for the other data sets and its neighboring values are not very high, so we regard the effect of this missing value as insignificant.

From Table 4 we see that for the bit fingerprint the addition of height 3 was the last one to make a significant difference (positive confidence interval). Also our nonsuperiority test for the bit fingerprint indicated that heights above height 3 were nonsuperior to staying at height 3. As for the count fingerprint, the addition of height 3 gave a significant difference (positive confidence interval). However, although it was statistically significant, that difference was not big enough to show up in our nonsuperiority test. At a significance level of 0.05, the addition of more signature heights above height 2 was found to be nonsuperior for the count fingerprint according to our nonsuperiority threshold. When these two nonsuperior height combinations for the count and bit fingerprints were used, no significant difference in performance between them could be seen (Wilcox $p = 0.81$).

On the basis of these results, a test to see the effect of removing the lower heights was set up using a grid search with tested $C = (0.001, 0.01, 0.1)$ and $\gamma = (1, 10, 100)$ and heights up to 3, where the lowest heights were removed one at a time. A series of noninferiority tests was performed, with an average difference of 0.01 in AUC once again considered relevant. The AUC values with confidence intervals and the noninferiority p values are given in Table 5. On the basis of the 95% confidence interval, we see no significant effect in removing height 0 for either the bit fingerprint or the count fingerprint, and at a significance level of 0.05 and our selected noninferiority threshold, keeping height 0 was nonsuperior to removing it.

DISCUSSION

On average, the best C and γ combination is the same for both the count and bit fingerprints, as can be seen in Figure 3. However, this does not necessarily mean that a grid search is not worth doing. In some cases another combination might be better.

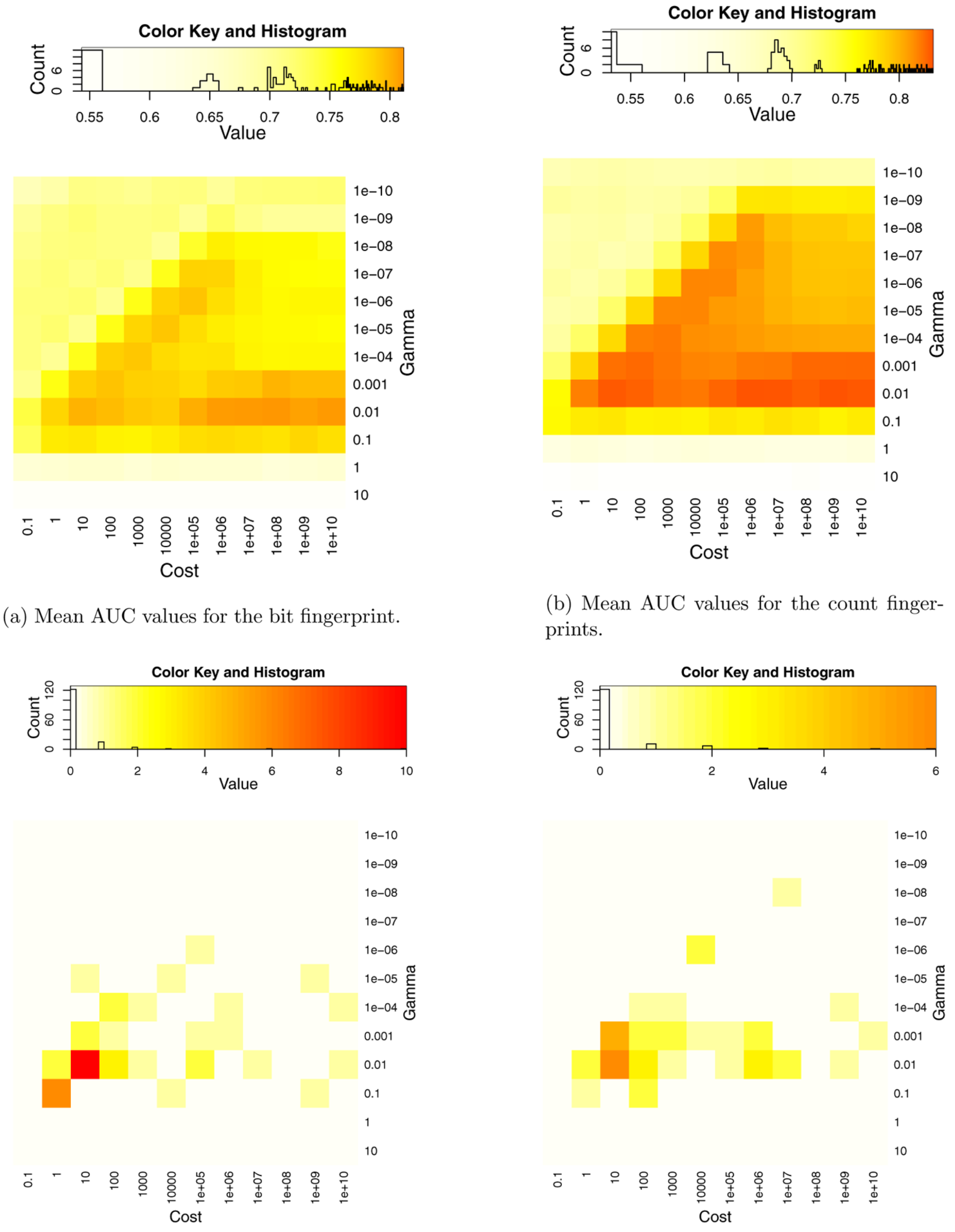
Figure 3 indicates that $C = 10$ and $\gamma = 0.01$ gives the highest AUC values but also that C values around 1×10^6 give decent values for γ around 0.01. In fact, there seems to be a γ interval where C almost does not matter. Thus, picking C and γ in the area around C from 1 to 100 and γ from 0.1 to 0.001 will most probably give a reasonably good result. A larger span might be worth trying if computational power is available, but we do not recommend C values higher than 1000, as all of the missing data in our experiments were generated with $C > 1000$. Thus, high C might be coupled with exceptionally long run times.

On the basis of our nonsuperiority tests for signature heights, different heights were chosen for the count and bit fingerprints. For the count fingerprint, a lower height resulted in performance similar to that for the bit fingerprint with a greater height.

It is worth noting that we did get a significant effect when height 3 was added for the count fingerprint (95% confidence interval between +0.001 and +0.008), but this increase in AUC was lower than the cutoff of 0.01 that was chosen for the nonsuperiority test. In regard to heights 4 and 5, our study does not indicate that adding them gives any benefits.

If reasons exist for removing the lower heights from the model, such as for visualization purposes, we see no relevant difference in performance when height 0 is removed. However, when greater heights are removed the performance goes down.

It might be tempting to take the AUC values in Table 4 and 5 as a measure of classification performance for the method, but it is important to remember that these values suffer from a positive bias since they are based on the same data set that was the basis for the selection of the maximally performing C and γ, and thus, they should be used only to study the effect of varying the signature height.



(a) Mean AUC values for the bit fingerprint.

(b) Mean AUC values for the count fingerprints.

(c) Number of times each combination of C and γ performed the best for the different heights and datasets using the bit fingerprint.

(d) Number of times each combination of C and γ performed the best for the different heights and datasets using the count fingerprint.

Figure 3. Visualization of the AUC values resulting from the different data sets and heights in the form of heat maps. Panels (a) and (b) show the average AUC values using a logarithmic color scale, and panels (c) and (d) show the number of times each combination of C and γ performed the best. All four heat maps show that the combination C = 10 and γ = 0.01 gives the highest AUC values.

We note that in Figure 2 one data set (CPDB) consistently performs worse. The Supporting Information shows that this data set has a very similar sweet spot for C and γ as the rest of the data sets. This further strengthens the interpretation that even with differently performing data sets our sweet spot still holds.

Studying the heat maps in the Supporting Information shows that cost becomes less important at greater heights. The cost determines the punishment for points that end up on the wrong side of the decision threshold. As the height increases, it seems to be easier to reach the same level of separation as with lower heights. However, we do not see that separation becomes better after height 2 or 3, only easier to find. Furthermore, we repeatedly see that the same sweet spot gives rise to the best performance. Hence, we recommend the choice of that sweet spot and staying at lower heights, and if computation time is available, a grid search around the sweet spot can be done instead of adding more heights.

■ CONCLUSIONS

Having a good prior knowledge of the impact model parameters can have on model performance in general can

Table 4. Wilcox Tests Performed on the Combination of C and γ Giving the Maximal AUC Value in Order To Study the Effect of Adding Signature Heights[a]

	AUC	95% CI for AUC	p value for nonsuperiority test
		Bit Fingerprint	
height 0	0.636	(0.605, 0.667)	
adding height 1	+0.193	(0.145, 0.247)	1.0
adding height 2	+0.019	(0.005, 0.028)	0.8516
adding height 3	+0.007	(0.002, 0.012)	0.1094
adding height 4	−0.001	(−0.004, 0.006)	0.0156
adding height 5	+0.001	(−0.006, 0.005)	0.0078
		Count Fingerprint	
height 0	0.769	(0.678, 0.844)	
adding height 1	+0.093	(0.042, 0.154)	1.0
adding height 2	+0.006	(−0.006, 0.016)	0.3438
adding height 3	+0.004	(0.001, 0.008)	0.0078
adding height 4	+0.001	(−0.004, 0.005)	0.0078
adding height 5	+0.002	(−0.001, 0.006)	0.0078

[a] The confidence interval (CI) is a nonparametric confidence interval calculated by the wilcox.test function in R. Nonsuperiority tests were performed using Wilcox tests looking at the change in AUC when another height is added using $H0$: ΔAUC > 0.01. At a significance level of 0.05, this would mean that for the bit fingerprint the addition of signature height 4 was nonsuperior to staying at height 3. For the count fingerprint, the addition of height 3 was nonsuperior to staying at height 2. After height 0 the AUC columns show the AUC difference from the previous height.

Table 5. AUCs and Confidence Intervals from a Grid Search with C = (0.001, 0.01, 0.1) and γ = (1, 10, 100) in Order To Study the Effect of Removing Lower Signature Heights[a]

	AUC	95% CI for AUC	p value for nonsuperiority test
		Bit Fingerprint	
height 0 to 3	0.867	(0.786, 0.911)	
removing height 0	+0.001	(−0.005, 0.008)	0.0078
removing height 0 to 1	−0.008	(−0.018, 0.000)	0.3438
removing height 0 to 2	−0.015	(−0.031, 0.002)	0.8125
		Count Fingerprint	
height 0 to 3	0.867	(0.790, 0.915)	
removing height 0	−0.001	(−0.005, 0.005)	0.0078
removing height 0 to 1	−0.013	(−0.025, 0.000)	0.7656
removing height 0 to 2	−0.022	(−0.035, −0.006)	0.9609

[a] The confidence interval (CI) is a nonparametric confidence interval calculated by the wilcox.test function in R. Non-inferiority tests were performed using Wilcox tests looking at the change in AUC compared with heights 0 to 3 when the lowest heights are removed using $H0$: ΔAUC > −0.01. Using a significance level of 0.05, this would mean that removing height 0 was noninferior to using the heights 0 to 3 by the noninferiority threshold value of a 0.01 difference in AUC. After height 0 the AUC columns show the AUC difference from the previous height.

reduce time and ease model building. On the basis of our threshold for what we judge to be a relevant difference in AUC, we recommend the parameter set of height 0 to 2 for the count version of the signature fingerprint and height 0 to 3 for the bit version. These are in combination with a support vector machine using C = 10 or a value chosen by a grid search between 1 to 100 and γ = 0.01 or a value chosen by grid search between 0.001 to 0.1, which provide a good starting point in the trade-off between speed and performance. However, we do see a statistically significant increase when height 3 is added for the count fingerprint. Thus, when data sets are small and execution time is not a problem, adding height 3 for the count fingerprint and using a grid search covering greater spans for C and γ might give a somewhat better result. Our data indicate that height 0 can be removed without causing any relevant degradation in model performance. Moreover, we see no benefit in using additional heights beyond height 3.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Heat maps for all heights and data sets. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: jonathan.alvarsson@farmbio.uu.se.

Notes

The authors declare the following competing financial interest(s): J.E.S.W., O.S., and M.E. hold shares in Genetta Soft AB, a Swedish incorporated company.

■ ACKNOWLEDGMENTS

The computations were performed on resources provided by SNIC through the Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) under Project p2010047. The authors thank Wesley Schaal for proofreading. This work was supported by the Swedish Strategic Research Program eSSENCE, the Swedish Council for Working Life and Social Research (FAS) (Grant 2012-0073), the Swedish VR (K2012-62X-22053-01-3), and the Swedish Foundation for Strategic Research (Grant RBa08-0114).

■ REFERENCES

(1) Hansch, C. Quantitative approach to biochemical structure−activity relationships. *Acc. Chem. Res.* **1969**, *2*, 232−239.

(2) Faulon, J.-L.; Visco, D. P.; Pophale, R. S. The signature molecular descriptor. 1. Using extended valence sequences in QSAR and QSPR studies. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 707−720.

(3) Faulon, J.-L.; Churchwell, C. J.; Visco, D. P. The signature molecular descriptor. 2. Enumerating molecules from their extended valence sequences. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 721−734.

(4) Faulon, J.-L.; Collins, M. J.; Carr, R. D. The signature molecular descriptor. 4. Canonizing molecules using extended valence sequences. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 427−436.

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

(6) Smola, A. J.; Schölkopf, B. A tutorial on support vector regression. *Stat. Comput.* **2004**, *14*, 199−222.

(7) Vanii, K.; Moreno, P.; Truszkowski, A.; Ertl, P.; Steinbeck, C. Natural product-likeness score revisited: An open-source, open-data implementation. *BMC Bioinf.* **2012**, *13*, 106.

(8) Carlsson, L.; Helgee, E. A.; Boyer, S. Interpretation of nonlinear QSAR models applied to Ames mutagenicity data. *J. Chem. Inf. Model.* **2009**, *49*, 2551−2558.

(9) Spjuth, O.; Eklund, M.; Ahlberg Helgee, E.; Boyer, S.; Carlsson, L. Integrated decision support for assessing chemical liabilities. *J. Chem. Inf. Model.* **2011**, *51*, 1840−1847.

(10) Spjuth, O.; Carlsson, L.; Alvarsson, J.; Georgiev, V.; Willighagen, E.; Eklund, M. Open source drug discovery with Bioclipse. *Curr. Top. Med. Chem.* **2012**, *12*, 1980−1986.

(11) Chen, H.; Carlsson, L.; Eriksson, M.; Varkonyi, P.; Norinder, U.; Nilsson, I. Beyond the scope of Free-Wilson analysis: Building

interpretable QSAR models with machine learning algorithms. *J. Chem. Inf. Model.* **2013**, *53*, 1324−1336.

(12) Weis, D. C.; Visco, D. P. Computer-aided molecular design using the Signature molecular descriptor: Application to solvent selection. *Comput. Chem. Eng.* **2010**, *34*, 1018−1029.

(13) Rostkowski, M.; Spjuth, O.; Rydberg, P. WhichCyp: Prediction of cytochromes P450 inhibition. *Bioinformatics* **2013**, *29*, 2051−2052.

(14) Lapins, M.; Worachartcheewan, A.; Spjuth, O.; Georgiev, V.; Prachayasittikul, V.; Nantasenamat, C.; Wikberg, J. E. A Unified Proteochemometric Model for Prediction of Inhibition of Cytochrome P450 Isoforms. *PloS One* **2013**, *8*, No. e66566.

(15) Norinder, U.; Ek, M. E. QSAR investigation of NaV1.7 active compounds using the SVM/Signature approach and the Bioclipse modeling platform. *Bioorg. Med. Chem. Lett.* **2013**, *23*, 261−263.

(16) Norinder, U.; Carlsson, L.; Boyer, S.; Eklund, M. Introducing Conformal Prediction in Predictive Modeling. A Transparent and Flexible Alternative to Applicability Domain Determination. *J. Chem. Inf. Model.* **2014**, *54*, 1596−1603.

(17) Alvarsson, J.; Eklund, M.; Engkvist, O.; Spjuth, O.; Carlsson, L.; Wikberg, J. E. S.; Noeske, T. Ligand-Based Target Prediction with Signature Fingerprints. *J. Chem. Inf. Model.* **2014**, DOI: 10.1021/ci500361u.

(18) Eklund, M.; Norinder, U.; Boyer, S.; Carlsson, L. *Artificial Intelligence Applications and Innovations*; Springer: New York, 2012; pp 166−175.

(19) Eklund, M.; Norinder, U.; Boyer, S.; Carlsson, L. The application of conformal prediction to the drug discovery process. *Ann. Math. Artif. Intell.* **2013**, 1−16.

(20) Hsu, C.-W.; Chang, C.-C.; Lin, C.-J. A Practical Guide to Support Vector Classification. 2010; http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf (accessed June 11, 2014).

(21) Steinbeck, C.; Han, Y.; Kuhn, S.; Horlacher, O.; Luttmann, E.; Willighagen, E. The Chemistry Development Kit (CDK): An open-source Java library for chemo- and bioinformatics. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 493−500.

(22) Steinbeck, C.; Hoppe, C.; Kuhn, S.; Floris, M.; Guha, R.; Willighagen, E. L. Recent Developments of the Chemistry Development Kit (CDK)—An Open-Source Java Library for Chemo- and Bioinformatics. *Curr. Pharm. Des.* **2006**, *12*, 2111−2120.

(23) Eklund, M.; Norinder, U.; Boyer, S.; Carlsson, L. Benchmarking Variable Selection in QSAR. *Mol. Inf.* **2012**, *31*, 173−179.

(24) Bruce, C. L.; Melville, J. L.; Pickett, S. D.; Hirst, J. D. Contemporary QSAR Classifiers Compared. *J. Chem. Inf. Model.* **2007**, *47*, 219−227.

(25) Chavatte, P.; Yous, S.; Marot, C.; Baurin, N.; Lesieur, D. Three-Dimensional Quantitative Structure−Activity Relationships of Cyclooxygenase-2 (COX-2) Inhibitors: A Comparative Molecular Field Analysis. *J. Med. Chem.* **2001**, *44*, 3223−3230.

(26) Sutherland, J. J.; Weaver, D. F. Three-dimensional quantitative structure−activity and structure−selectivity relationships of dihydrofolate reductase inhibitors. *J. Comput.-Aided Mol. Des.* **2004**, *18*, 309−331.

(27) Gold, L. S.; Manley, N. B.; Slone, T. H.; Rohrbach, L. Supplement to the Carcinogenic Potency Database (CPDB): Results of animal bioassays published in the general literature in 1993 to 1994 and by the National Toxicology Program in 1995 to 1996. *Environ. Health Perspect.* **1999**, *107* (Suppl.), 527−600.

(28) Gold, L. S.; Manley, N. B.; Slone, T. H.; Ward, J. M. Compendium of chemical carcinogens by target organ: Results of chronic bioassays in rats, mice, hamsters, dogs, and monkeys. *Toxicol. Pathol.* **2001**, *29*, 639−652.

(29) Gold, L. S.; Manley, N. B.; Slone, T. H.; Rohrbach, L.; Garfinkel, G. B. Supplement to the Carcinogenic Potency Database (CPDB): Results of animal bioassays published in the general literature through 1997 and by the National Toxicology Program in 1997−1998. *Toxicol. Sci.* **2005**, *85*, 747−808.

(30) Russom, C. L.; Bradbury, S. P.; Broderius, S. J.; Hammermeister, D. E.; Drummond, R. A. Predicting Modes of Toxic Action from Chemical Structure: Acute Toxicity in the Fathead Minnow (*Pimephales promelas*). *Environ. Toxicol. Chem.* **1997**, *16*, 948−967.

(31) Matthews, E. J.; Kruhlak, N. L.; Benz, R. D.; Contrera, J. F. Assessment of the health effects of chemicals in humans: I. QSAR estimation of the maximum recommended therapeutic dose (MRTD) and no effect level (NOEL) of organic chemicals based on clinical trial data. *Curr. Drug Discovery Technol.* **2004**, *1*, 61−76.

(32) Hansen, K.; Mika, S.; Schroeter, T.; Sutter, A.; ter Laak, A.; Steger-Hartmann, T.; Heinrich, N.; Müller, K.-R. Benchmark data set for in silico prediction of Ames mutagenicity. *J. Chem. Inf. Model.* **2009**, *49*, 2077−2081.

(33) ftp://ftp.ics.uci.edu/pub/baldig/learning/nci/gi50/ (accessed May 14, 2014).

(34) Bradley, A. P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit.* **1997**, *30*, 1145−1159.

(35) Schumi, J.; Wittes, J. T. Through the looking glass: Understanding non-inferiority. *Trials* **2011**, *12*, 106.

(36) R Development Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2008.

(37) Dimitriadou, E.; Hornik, K.; Leisch, F.; Meyer, D.; Weingessel, A. *e1071: Misc Functions of the Department of Statistics (e1071)*; TU Wien: Vienna, Austria, 2011; R package, version 1.6.

(38) Robin, X.; Turck, N.; Hainard, A.; Tiberti, N.; Lisacek, F.; Sanchez, J.-C.; Müller, M. pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinf.* **2011**, *12*, No. 77.

(39) Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*; Springer: New York, 2009.

(40) Warnes, G. R.; Bolker, B.; Bonebakker, L.; Gentleman, R.; Liaw, W. H. A.; Lumley, T.; Maechler, M.; Magnusson, A.; Moeller, S.; Schwartz, M.; Venables, B. *gplots: Various R Programming Tools for Plotting Data*; 2011; R package, version 2.10.1.