Interpretation of Nonlinear QSAR Models Applied to Ames Mutagenicity Data

Lars Carlsson,* Ernst Ahlberg Helgee, and Scott Boyer

Safety Assessment, AstraZeneca Research & Development, 43183 Mölndal, Sweden

A method for local interpretation of QSAR models is presented and applied to an Ames mutagenicity data set. In the work presented, local interpretation of Support Vector Machine and Random Forest models is achieved by retrieving the variable corresponding to the largest component of the decision-function gradient at any point in the model. This contribution to the model is the variable that is regarded as having the most importance at that particular point in the model. The method described has been verified using two sets of simulated data and Ames mutagenicity data. This work indicates that it is possible to interpret nonlinear machine-learning methods. Comparison to an interpretable linear method is also presented.

1. INTRODUCTION

Quantitative-structure activity-relationship (QSAR) models are widely used as a way of approximating a functional relationship between a set of descriptors and a given endpoint. Model interpretation is very important, without it the possibility of modifying compounds is restricted. Often chemists resort to model algorithms that are easy to understand, such as linear models. Nonlinear models are commonly viewed as hard to interpret, and these model algorithms or models are sometimes referred to as black-box models.[1,2] Some phenomena will most likely be modeled best by nonlinear models, and there is a need for QSAR models that are easy to interpret.

For some model algorithms there are ways to extract important attributes, for example the **importance** in Random Forests, R.[3] However, these methods usually provide the globally most important attributes. This is sometimes of limited utility when deciding what modification to make to a particular molecule to change its behavior against a specific biological activity since it would only indicate what attributes are important in the overall model but not in that particular local model space. One could want to change a specific attribute that might not be very important when a complete training set of compounds is considered. This can be done by looking at the local behavior of the function described by the QSAR model. One way is to compute the gradient, the partial derivatives of each attribute, and view the gradient and the prediction as a local description of the function, a truncated Taylor expansion.

Previous work in the field has been primarily focused on the possibility of finding the most important variables globally. In ref 4 gradients are computed once in each variable, and then the contributions are added to each molecule in order to achieve the globally most important variables. In ref 5 however, they divide the global chemical space modeled into subspaces and use linear regression to model these smaller sets of data, and they then discuss the issue of interpretability for the data as a whole.

The remainder of this paper is organized as follows. In section 2 methods are presented describing the retrieval of the local behavior, both for analytical and discrete formulations of QSAR-model functions. In Section 3, two examples have been constructed using simulated data over a predefined grid; the method has also been applied to Ames mutagenicity data. The paper is concluded by the discussion and conclusions in Section 4.

2. METHOD

Any QSAR model is an approximation to a relation between biological activity and compound characteristics and can be viewed as a mathematical function. Different machine-learning methods have different ways of deriving these approximations. More detailed descriptions of the methods used in this work can be found in the references for Partial Least Squares[6] (PLS), Random Forests[7] (RF), and in Support Vector Machines[8] (SVM). In the scope of this work, it suffices to realize that some methods have simple explicit expressions, whereas other methods result in more complex expressions.

The gradient of a function in a point tells how this function behaves in the neighborhood to that point, and it can be calculated for any sufficiently smooth function. By computing the gradient of a QSAR model, a linear approximation of the model can be retrieved. This enables the exploration of chemical space in that local neighborhood, and by changing specific descriptors it is then possible to reduce problems or rather optimize a query compound based on a given objective.

In the current work only the descriptor that makes the largest change to the function will be considered, that is the descriptor corresponding to the largest component of the gradient. When looking at a prediction of a specific class the descriptor that contributes the most is the interesting one. Alternative ways to identify important descriptors could be to take the *n* largest components or to use a cutoff value that defines the components to be picked or one could look at how a decrease could be obtained. This can be adjusted to suit the specific application.

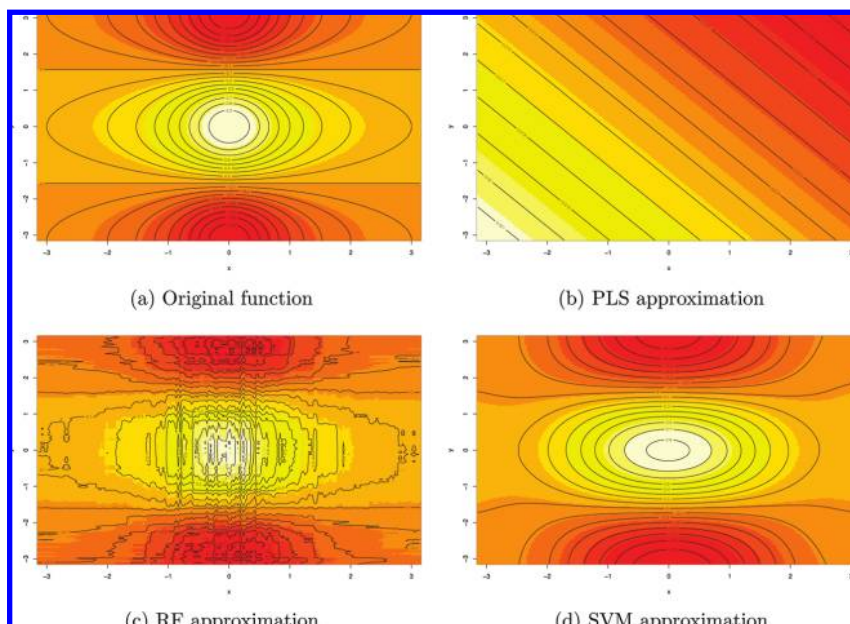* Corresponding author e-mail: lars.a.carlsson@astrazeneca.com.

Figure 1. Contour plots of the function for *Data set I*.
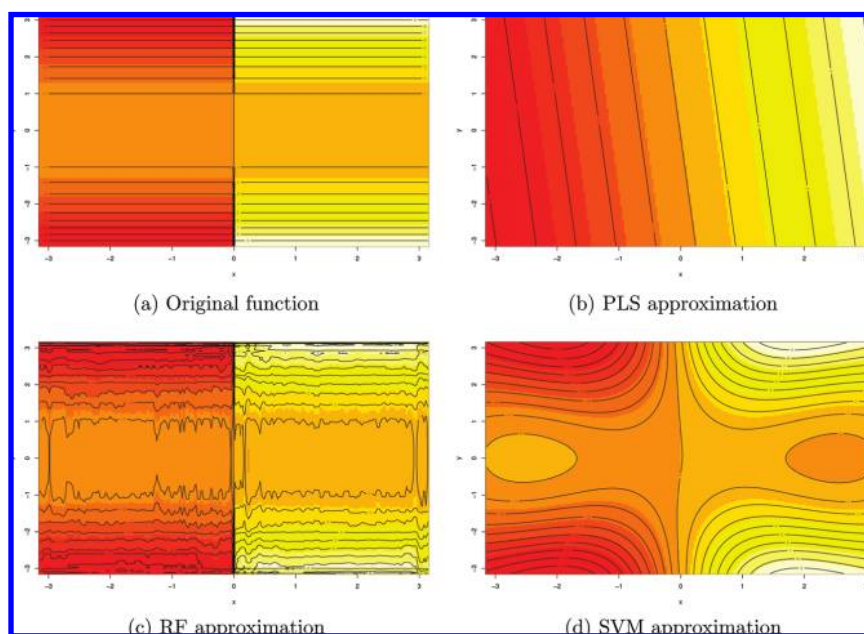

Figure 2. Contour plots of the function for *Data set II*.

The gradient of the function can be computed in different ways. Some machine-learning algorithms have decision functions that can be expressed analytically, for example SVM. In this case it is straightforward to derive an analytical expression of the function. For tree-based methods, like RF, there is no simple mathematical expression describing the decision function since it is composed by many piecewise constant functions assembled in an ensemble of trees. It is possible to derive it, but it would be very time-consuming. Instead, a discrete gradient of the function can be computed.

Consider a set of compounds and a given activity $\{\mathbf{x}_k, y_k\}$, the descriptors for the kth compound is denoted $\mathbf{x}_k$ and the corresponding activity y_k, which can be either discrete or continuous. These compounds are used to model the relationship between descriptors and activity. The relationship

can be regarded as a function $y = f(\mathbf{x})$. The decision function for the binary classifier, C-SVC model, built using libSVM[9] can be expressed as

$$f_{SVM} = \sum_{i=1}^{l} \alpha_i K(\mathbf{x}_i, \mathbf{x}) + \rho \qquad (1)$$

where $\mathbf{x}$ is the descriptor. This notation is not to be confused with the one that the authors of libSVM are using. It relates to the output in model files of the C-SVC method. The subscript i is an enumerator of the l support vectors forming the model, where the ith support vector is denoted $\mathbf{x}_i$ and the corresponding weight is α_i. The bias term is ρ. If f_{SVM} is greater than zero the call is for one class, otherwise it is for the other class. Everything is constant in the decision function except the kernel function K, hence deriving the gradient of
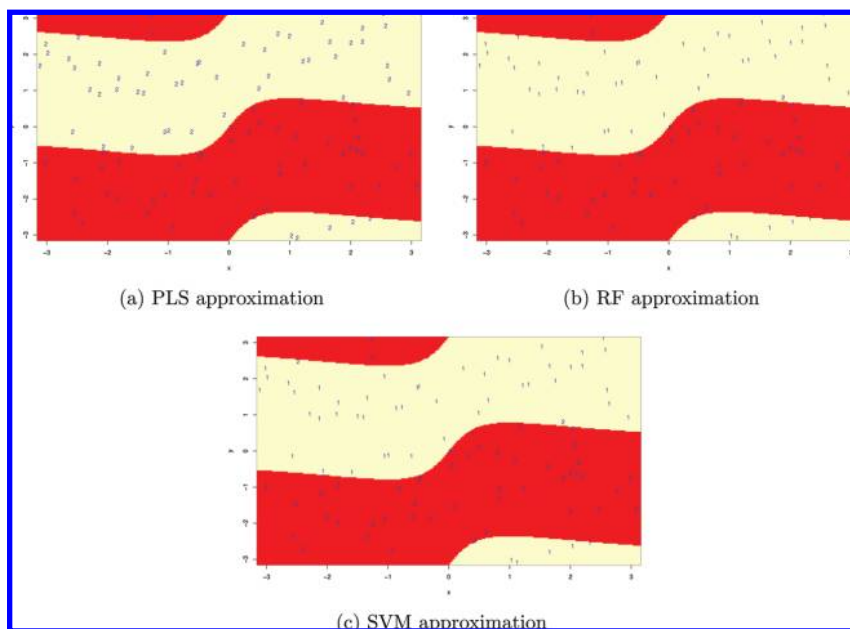
Figure 3. The most significant component of the gradient is plotted at each test point. This is based on the gradient computed for the PLS, RF, and SVM models with the first seed and a training set size of 800 data points. The underlying contour plot represents the correct areas of significance, yellow indicating component 1 and red indicating component 2.
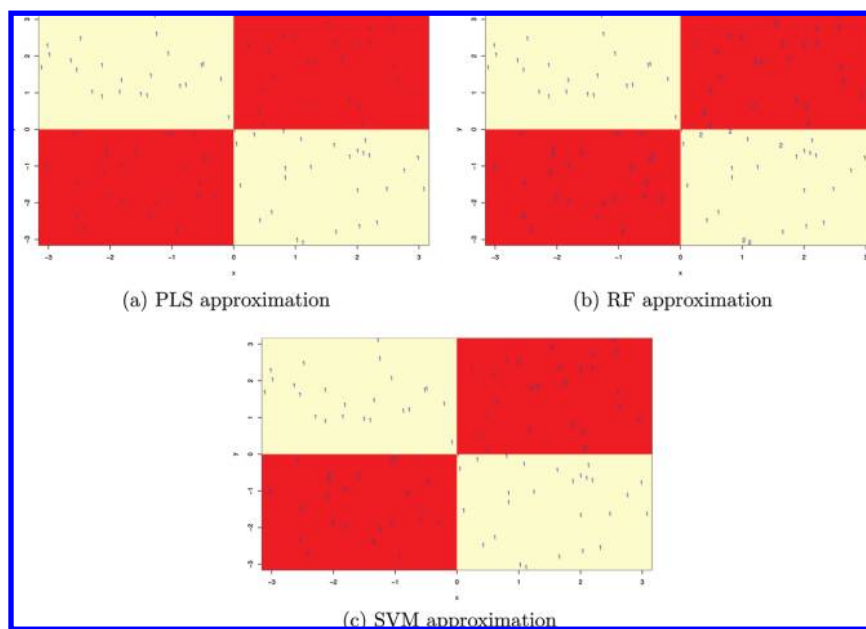

Figure 4. The most significant component of the gradient is plotted at each test point. This is based on the gradient computed for the PLS, RF, and SVM models with the first seed and a training set size of 800 data points. The underlying contour plot represents the correct areas of significance, yellow indicating component 1 and red indicating component 2.

f_{SVM} amounts to computing the partial derivatives of the kernel function. Throughout this study the radial-basis function will be used as the kernel function for the SVM models that are built

$$K(\mathbf{x}_i, \mathbf{x}) = e^{-\gamma(\mathbf{x}_i - \mathbf{x})(\mathbf{x}_i \ - \ \mathbf{x})^T} \qquad (2)$$

where γ is a constant specified in the libSVM model file. The jth partial derivative of the kernel function is

$$\frac{\partial K}{\partial x_j} = 2\gamma |x_i \ - \ x|_j e^{-\gamma(\mathbf{x}_i - \mathbf{x})(\mathbf{x}_i \ - \ \mathbf{x})^T} \qquad (3)$$

The jth component of the gradient of the decision function is

$$\nabla f_{SVM,j} = \sum_i^l 2\alpha_i \gamma |x_i \ - \ x|_j e^{-\gamma(\mathbf{x}_i - \mathbf{x})(\mathbf{x}_i \ - \ \mathbf{x})^T} \qquad (4)$$

Furthermore, the RF[3] and the PLS[10] packages in R will also be used to generate models. In the case of RF models, in general, there is no easy way of obtaining an analytical expression of the decision function. Instead, for this model and similar models one can compute the jth component of the discrete gradient

$$\frac{Df}{Dx_j} = \frac{\beta_1 f'(\mathbf{x} + h_j) + \beta_2 f'(\mathbf{x}) + \beta_1 f'(\mathbf{x} - h_j)}{2\beta_1 + \beta_2} \quad (5)$$

where β_1 and β_2 are smoothing coefficients. The step size in the *j*-direction in attribute space is h_j, and the corresponding second-order accurate partial derivative is $f'_j = (f(\mathbf{x} + h_j) - f(\mathbf{x} - h_j))/2h_j$. A smoothing has been introduced since many tree-based methods are piecewise constant in attribute space, and this can make the decision function a bit rough.

To evaluate the method of interpreting QSAR models by computing the largest component of the gradient, a case where the characteristics of compounds has shown to be valid and useful is needed. The prediction of Ames mutagenicity is such an example, and the SMARTS[11] defined as toxicophores[12] are specific in the sense that they are easy to visualize and to understand. By using attributes that can capture the same type of information as the SMARTS although generated independently of the SMARTS a valid comparison can be made.

The attribute to use when applying the current method to Ames mutagenicity data is the signature descriptor.[13,14] This descriptor describes a compound by a set of strings. Each string is a canonical representation of an atom and its neighboring atoms. The number of neighbors included can

Table 1. Average Mean-Square Error (MSE) and the Accuracy for the Gradient (GA) with Different Step Sizes, Respectively, Computed for All Five Different Seeds for *Data Set I* at the Various Training Set Sizes

	PLS			RF			SVM		
size	MSE	GA (h)	GA (4h)	MSE	GA (h)	GA (4h)	MSE	GA (h)	GA (4h)
100	$1.1e{-}1$	52	52	$3.3e{-}2$	68	79	$8.5e{-}3$	90	90
200	$1.1e{-}1$	50	50	$2.0e{-}2$	71	85	$5.9e{-}3$	90	90
400	$1.1e{-}1$	49	49	$7.0e{-}3$	77	90	$4.0e{-}3$	92	92
800	$1.1e{-}1$	49	49	$3.3e{-}3$	81	94	$2.5e{-}3$	91	92

Table 2. Average Mean-Square Error (MSE) and the Accuracy for the Gradient (GA) with Different Step Sizes, Respectively, Computed for All Five Different Seeds for *Data Set II* at the Various Training Set Sizes

	PLS			RF			SVM		
size	MSE	GA (h)	GA (4h)	MSE	GA (h)	GA (4h)	MSE	GA (h)	GA (4h)
100	$1.2e1$	50	50	$4.3e0$	67	70	$3.8e0$	80	79
200	$1.1e1$	50	50	$2.4e0$	70	77	$2.5e0$	83	83
400	$1.1e1$	50	50	$1.0e0$	71	83	$2.2e0$	86	86
800	$1.1e1$	50	50	$4.0e{-}1$	78	87	$2.1e0$	87	87

Table 3. Confusion Matrix for All Models on *Test Set I*

	pred. Ames mutagen	pred. Ames nonmutagen
	PLS	
exp. Ames mutagen	251(30%)	101(12%)
exp. Ames nonmutagen	284(34%)	196(24%)
	RF	
exp. Ames mutagen	249(30%)	103(12%)
exp. Ames nonmutagen	187(23%)	293(35%)
	SVM	
exp. Ames mutagen	259(31%)	93(11%)
exp. Ames nonmutagen	193(23%)	287(35%)
	SMARTS	
exp. Ames mutagen	235(28%)	117(14%)
exp. Ames nonmutagen	200(24%)	280(34%)

Table 4. Confusion Matrix for All Models on *Test Set II*

	pred. Ames mutagen	pred. Ames nonmutagen
	PLS	
exp. Ames mutagen	337(38%)	213(24%)
exp. Ames nonmutagen	157(18%)	172(20%)
	RF	
exp. Ames mutagen	453(52%)	97(11%)
exp. Ames nonmutagen	66(7%)	263(30%)
	SVM	
exp. Ames mutagen	437(50%)	113(13%)
exp. Ames nonmutagen	81(9%)	248(28%)
	SMARTS	
exp. Ames mutagen	439(50%)	111(13%)
exp. Ames nonmutagen	92(10%)	237(27%)

Table 5. Generated SMARTS and the Statistics Corresponding to the Toxicophores Definitions for the Most Significant Contributions in Table 9

data set:compound number	positive occurrences	total occurrences	accuracy	p-value	SMARTS from fragments
1:32	988	1342	0.74	2.1e-42	Nc(c)c
1:34	5	5	1.00	5.3e-02	cC(N)=C
1:35	1003	1734	0.58	3.3e-02	cc(c)C
1:250	1019	1284	0.79	2.5e-71	cc(c)c
1:302	988	1342	0.74	2.1e-42	cc(c)N
1:570	128	194	0.66	2.1e-03	Cn(c)c
1:570	45	56	0.80	9.4e-05	nc(c)-c
1:689	1019	1284	0.79	2.5e-71	cc(c)c
1:756	293	594	0.49	1.0e00	OC(C)C
1:796	87	109	0.80	1.0e-07	O=C(c)c
1:817	425	836	0.51	1.0e00	Oc(c)c
2:84	7	7	1.00	1.6e-02	$N = N{=}C$
2:508	293	594	0.49	1.0e00	OC(C)C
2:524	428	576	0.74	1.3e-20	nc(c)c
2:681	64	101	0.63	7.0e-02	OC(C)O

vary depending on the predefined distance threshold, called height. At height zero, the signature descriptor is an enumeration of atom types in a compound. For any other height, simple bond types are included. The number of signatures depends on the number of unique atoms and their environments. The signature descriptors are generated by the software package **translator**.[15]

3. RESULTS

This section describes a numerical experiment on a known relationship and the application to Ames mutagenicity data.

3.1. A Numerical Experiment on Simulated Data. The method has been applied to two separate data sets of simulated data, *Data set I* and *Data set II* and implemented in **R**.[16] Both data sets have a predefined relationship between attributes and responses. The method of predefined solutions has been successful in assessing methods in other scientific areas.[17] Attribute space is two-dimensional and identical for both data sets, and it spans a range from $-\pi$ to π, for each attribute. The data points in attribute space are uniformly distributed, and there are 200 points for each attribute. A fixed test set size of 100 data points has been used, and the training set size has been varied, 100, 200, 400, and 800 data points. The function defining the relationship for *Data set I* is $f_I = \cos x_2/(1 + x_1^2)$, Figure 1, and for *Data set II* the function is $f_{II} = x_2^2|x_1|/x_1$, Figure 2, where x_1 and x_2 are the variables in attribute space. For each data set, the relationship between the attributes and the responses is computed in every

Table 6. Compounds That Where Previously Predicted As Negative Where the Most Significant Contribution Has Been Removed
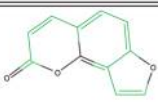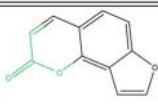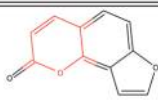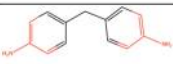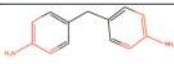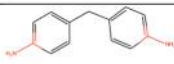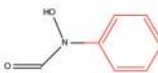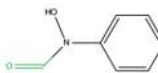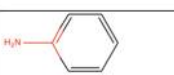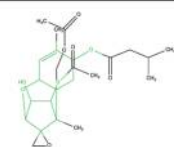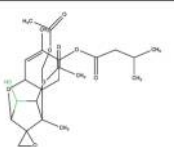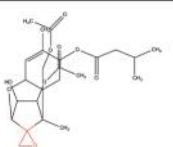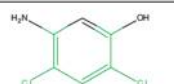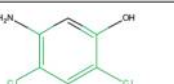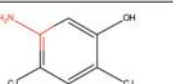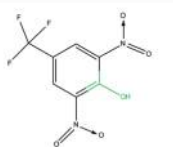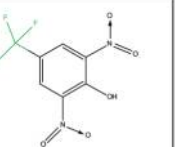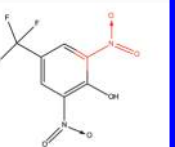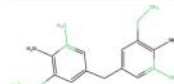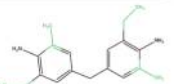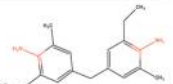
Dataset : Compound number	Significant SVM fragment	Significant RF fragment	SMARTS	DB POS:NEG occurrences
1:300				A2H 2560 0:3
1:473				A2H 4124 8:6
1:785				A2H 4636 0:5
2:49				A2H 2565 0:67

Table 7. Negative Compounds That Have Negative Predictions for SVM and RF but a Positive Prediction for the SMARTS Model

Dataset : Compound number	Activity	Significant SVM fragment	Significant RF fragment	SMARTS
1:3	NEG			
1:194	NEG			
1:356	NEG			
1:473	NEG			

data point. Then data points are picked randomly without replacement for the training and test sets, respectively. This is repeated with five different seeds for each training set size.

Three different machine-learning algorithms are applied to the simulated data. These methods are available in **R**, PLS (**pls**),[10] RF (**randomForest**),[3] and SVM (**e1071**).[18] Each model is trained to do regression predictions, and for every prediction the discrete gradient of the model function is computed as a metric for the most significant attribute and can be compared to the analytical gradient. The models were trained using their default settings, respectively. Examples of PLS, RF, and SVM approximating the function is given in Figures 1 and 2.

The roughness of the approximated functions might be an issue when looking at their gradients. To further investigate this a simple test was set up. A smoothing was introduced in eq 5 with $\beta_1 = 1$ and $\beta_2 = 2$. The step size was chosen to be identical in each attribute direction and set to a length where $h = 2\pi/(200 + 1)$. Another increased step size was also used, and it was four times as large as the

Table 8. Negative Compounds That Have Positive Predictions for SVM and RF but a Negative Prediction for the SMARTS Model
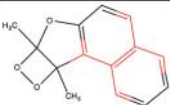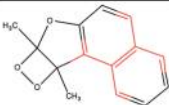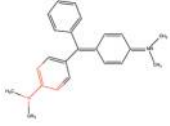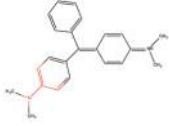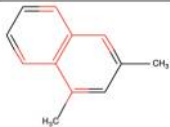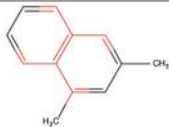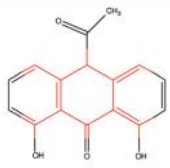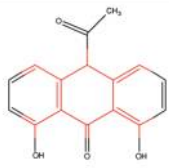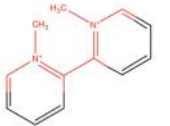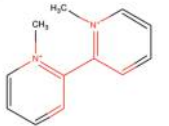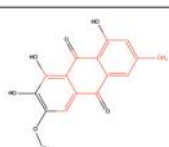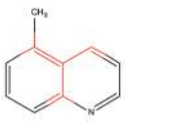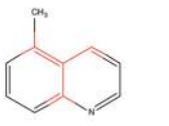
Dataset : Compound number	Activity	Significant SVM fragment	Significant RF fragment	SMARTS
1:51	NEG			-
1:112	NEG			-
2:501	NEG			-
2:511	NEG			-

Table 9. Positive Compounds That Have Positive Predictions for SVM and RF but a Negative Prediction for the SMARTS Model

Dataset : Compound number	Activity	Significant SVM fragment	Significant RF fragment	SMARTS
1:35	POS			-
1:570	POS			-
2:714	POS			-
2:736	POS			-
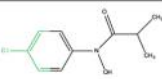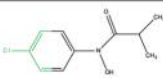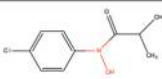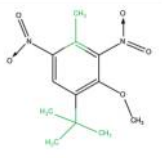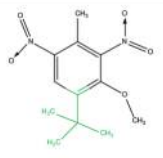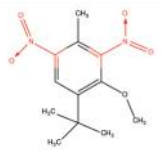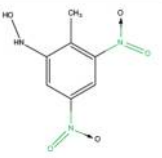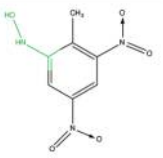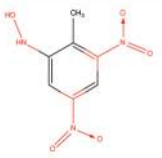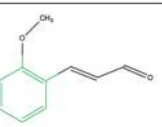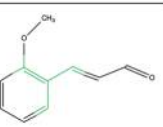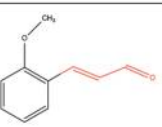
original one. A summary of the test set statistics for these methods is displayed in Tables 1 and 2.

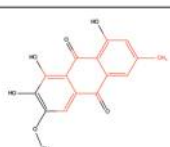Examples of the most significant components for *Data set I* are presented in Figure 3 and the corresponding plots for *Data set II* are presented in Figure 4.

3.2. Ames Mutagenicity Prediction. When working with Ames mutagenicity data, the focus has been on three different

data sets. The data sets do not describe the mechanisms of Ames mutagenicity in detail but rather whether the compounds in the data sets are mutagenic or not. The first data set has been used as a training set, *Training set*, and it is based on the data reported by Kazius et al.[12] The two remaining data sets were used as test sets. *Test set I* was obtained from the Chemical Carcinogenicity Research In-

QSAR Models Applied to Ames Mutagenicity Data

J. Chem. Inf. Model., Vol. 49, No. 11, 2009 **2557**

Table 10. Positive Compounds That Have Negative Predictions for SVM and RF but a Positive Prediction for the SMARTS Model

Dataset : Compound number	Activity	Significant SVM fragment	Significant RF fragment	SMARTS
1:12	POS			
1:74	POS			
1:223	POS			
2:182	POS			

formation System database,[19] and *Test set II* a separate collection was reported by Young et al.[20] All three data sets, described as SMILES strings, were processed through a series of filters starting with an AstraZeneca in-house tool, LEATH-ERFACE[21] to derive canonical tautomers. Chirality has been removed by simple text substitution, for example C[@@H] has been changed to C. Furthermore, inorganic and organo-metallic compounds have all been excluded using filters written based on the Openeye Toolkit.[22] The activity for a compound reported more than once was treated as reported by Kazius et al.[12] The data sets were then compared pairwise against each other to exclude duplicates. The highest priority was given *Training set* and then *Test set II* and then the lowest priority to *Test set I*.

Signature descriptors of height 1 were generated for *Training set*, and any descriptor that varied in less than 0.1% of the examples was excluded. This resulted in a training set of 4254 compounds of which 2366 are Ames mutagens. The total number of compounds in *Test set I* is 832 with 352 Ames mutagens, and for *Test set II* the corresponding numbers are 879 and 550. The same set of descriptors, a total of 192, was generated for the two test sets.

Four different models were built, a PLS, an RF, a SMARTS, and an SVM model. The SMARTS model was based on the 28 approved toxicophores defined by Kazius et al.[12] and implemented using the Openeye Toolkit.[22] Two SMARTS patterns were modified; sulfonate bonded carbon [C,c]OS((=O)=O)O!@[c,C] and [c,C]S((=O)=O)O!@[c,C] and α- and β- unsaturated aldehyde O=[CH]C=[C,O] and not [$(O =[CH]C([N, O, S]) = C); $ (O=[CH]C=C[N,O,S]); $(O=[CH]C=Ca)]. The SVM model was picked from a 5-fold cross-validation procedure where the parameters γ and C were optimized using the grid-search method included in libSVM.[9]

Some statistics for the models are presented in Tables 3 and 4. The overall accuracies for *Test set I* were 66% for

the SVM model, 65% for the RF model, 62% for the SMARTS model, and 54% for the PLS model. The corresponding numbers for *Test set II* were 78% for the SVM model, 81% for the RF model, 77% for the SMARTS model, and 58% for the PLS model. Subsets of the predicted compounds are shown in Tables 7−10. In this case no smoothing was used, hence $\beta_1 = 1$ and $\beta_2 = 0$ in eq 5. Furthermore, the step size was set to unity in each attribute direction. This was motivated by the use of the signatures as they only can take on integer values.

To further test the proposed method SMARTS were generated for the significant parts of some of the compounds in Table 9. Statistics corresponding to the analysis presented by Kazius et al.[12] were computed, and the numbers are presented in Table 5. The *p*-value used in this work is defined as $\sum_{k=r}^{n} \binom{n}{k} p^k (1 - p)^{(n-k)}$, where n is the total number of occurrences of the specific structure in the data set, and r is the number of occurrences in compounds being positive and $p = r/n$.

Table 6 displays compounds where the most significant fragments have been removed or replaced, and the remainder of each compound have been predicted and evaluated by the proposed method. The compounds were also checked against a database provided by MultiCASE.[23]

4. DISCUSSION AND CONCLUSIONS

The behavior of the underlying relationship can be controlled by using simulated data. Thus, it gives an opportunity to study errors in the approximated QSAR models. Furthermore, it can show differences between different modeling algorithms. Tables 1 and 2 show the accuracies for the prediction of the most significant attribute. By using the simulated data it is possible to see that the proposed method works in the case of nonlinear algorithms like the SVM and the RF. The choice of an appropriate step

size for computing the discrete gradient has not been established, and further investigations are needed. For PLS, however, the results are poor which is not surprising since this method can only describe linear relationships. A linear method cannot accurately describe nonlinear data, and in terms of interpretability linear methods are of less value when applied to nonlinear relationships as seen in Figures 1 and 2.

The PLS model is unable to determine the locally most significant contribution to the prediction since it is a linear model and the gradient is constant throughout attribute space. For the RF model there is a dramatic increase in performance for the prediction of the most significant component as the step size, $h = 2\pi/(200 + 1)$, is increased. The SVM model does not appear to be sensitive to this change in step size. For both step sizes the smoothed gradient operator (5) with $\beta_1 = 1$ and $\beta_2 = 2$ have been applied. The difference in accuracy for a varying step size might be a problem when the gradients are computed. The original step size used for the simulated data corresponds to how well the function is resolved or sampled in each attribute direction. There is no guarantee that a step size chosen this way would give accurate results. In the case of the Ames mutagenicity data, both SVM and RF give similar results when the step size is chosen based on the resolution in the attributes.

By computing the gradient of a QSAR model it is indeed possible to describe the behavior in a local neighborhood to a prediction. When applied to mutagenicity data described by the signature descriptor this method shows very nice correspondence to previously published toxicophores. The model performance is slightly better with the proposed method than by using a model based on the toxicophores. It can also be seen that the local behavior of a nonlinear machine-learning method is of great importance. As illustrated in Table 5 there are both significant and nonsignificant fragments in terms of the analysis Kazius et al. conducts.[12] If the fragments from Table 5 are described as SMARTS and used as proposed by Kazius et al., important contributions will be overseen since not all fragments are accurate enough. This will reduce the possibility to understand the underlying chemistry. If the method proposed in this paper is used, it will instead enhance the understanding of the compound specific problem. This is also in good agreement with the results presented in Table 6, where fragments that contribute significantly to a negative prediction have been removed or replaced and the prediction changes for some of the compounds. It is also interesting to note that a compound that contains a toxicophore does not have to be positive in the Ames mutagenicity test and that a nonlinear method can make a correct prediction, as can be seen in Tables 6 and 7. Both the proposed method and the method described by Kazius et al. are approximations based on a different hypothesis, and hence there will always be cases where one method predicts a single compound correct when the other fails. Instead of having a set of substructures that in general can explain a specific behavior, like Ames mutagenicity, it might be preferable to have an accurate approximation of this behavior and to be able to interpret the individual predictions. If interpretation is necessary there is no need to use linear methods only. Variable importance proposed by the use of other methods can be misleading. It only shows the global behavior, not the very important local

behavior that really tells a chemist what the problem is. The results on the simulated data clearly shows this.

This method can be used with any type of descriptors that allows for the gradient to be computed. The model results may be different, but they would still be interpretable. The application to Ames mutagenicity data used structural descriptors, signatures, to allow for the comparison to the toxicophores by Kazius et al. The differences in overall accuracies are not great, but the models presented here performs slightly better than a SMARTS model based on the toxicophores.

Finally, additional information that the gradient gives might improve the method. In particular, more than a single component can be used. This is an area that needs further attention.

Supporting Information Available: Table, with predictions and the largest component of the gradient corresponding to the prediction, for each of the two test sets. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Fox, T.; Kriegl, J. M. *Curr. Top. Med. Chem* **2006**, *6*, 1579–1591.
(2) Helma, C.; Kazius, J. *Curr. Comput.-Aided Drug Des.* **2006**, *2*, 123–133.
(3) Liaw, A.; Wiener, M. Breiman and Cutler's random forests for classification and regression, version 4.5-18; 2006, The Comprehensive R Archive Network. http://cran.r-project.org/doc/packages/randomForest.pdf (accessed Jan 3, 2007).
(4) Franke, L.; Byvatov, E.; Werz, O.; Steinhilber, D.; Schneider, P.; Schneider, G. *J. Med. Chem.* **2005**, *48*, 6997–7004.
(5) Guha, R.; Dutta, D.; Jurs, P. C.; Chen, T. *J. Chem. Inf. Model.* **2006**, *46*, 1836–1847.
(6) Wold, S.; Ruhe, A.; Wold, H.; Dunn III, W. J. *SIAM J. Sci. Stat. Comput.* **1984**, *5*, 735–743.
(7) Breimann, L. *Machine Learning* **2001**, *45*, 5–32.
(8) Christianini, N.; Shawe-Taylor, J. *An Introduction to Support Vector Machines and other kernel-based learning methods*, 1st ed.; Cambridge University Press: Cambridge, United Kingdom, 2004.
(9) Chang, C.-C.; Lin, C.-J. LIBSVM: a library for support vector machines; 2001; Software available at http://www.csie.ntu.edu.tw/cjlin/libsvm.
(10) Wehrens, R.; Mevik, B.-H. Partial Least Squares Regression (PLSR) and Principal Component Regression (PCR), version 2.1.0; 2007, The Comprehensive R Archive Network. http://cran.r-project.org/doc/packages/pls.pdf (accessed Sep 11, 2007).
(11) Daylight Theory: SMARTS - A Language for Describing Molecular Patterns. http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html (accessed Jan 13, 2008).
(12) Kazius, J.; McGuire, R.; Bursi, R. *J. Med. Chem.* **2005**, *48*, 312–320.
(13) Faulon, J.-L.; Visco, D. P. J.; Pophale, R. S. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 707–720.
(14) Faulon, J.-L.; Churchwell, C. J. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 721–734.
(15) Faulon, J.-L. http://www.cs.sandia.gov/ jfaulon/QSAR/translator.tar.gz.
(16) R, version 2.5.0; 2007, The R Project for Statistical Computing. http://www.r-project.org (accessed Jan 3, 2007).
(17) Henshaw, W. D. *J. Comput. Phys.* **1994**, *113*, 13–25.
(18) Dimitriadou, E.; Hornik, K.; Leisch, F.; Meyer, D.; Weingessel, A. Misc Functions of the Department of Statistics (e1071), TU Wien, version 1.5−16; 2006, The Comprehensive R Archive Network. http://cran.r-project.org/doc/packages/e1071.pdf (accessed Jan 3, 2007).
(19) TOXNET - Chemical Carcinogenesis Research Information System, http://toxnet.nlm.nih.gov (accessed Nov 22, 2006).
(20) Young, S.; Gombar, V.; Emptage, M.; Cariello, N.; Lambert, C. *Chemom. Intell. Lab. Syst.* **2002**, *60*, 5–11.
(21) Kenny, P.; Sadowski, J. *Chemoinformatics Drug Discovery* **2005**, 271–285.
(22) Openeye Scientific Software. http://www.eyesopen.com (accessed Aug 30, 2005).
(23) MultiCASE Inc. http://www.multicase.com/products/prod0910.htm (accessed Jan 13, 2008).