Evaluation of Quantitative Structure−Activity Relationship Modeling Strategies: Local and Global Models

Ernst Ahlberg Helgee,*,† Lars Carlsson,† Scott Boyer,† and Ulf Norinder‡

Safety Assessment, AstraZeneca Research & Development, 43183 Mölndal, Sweden, and Medicinal Chemistry, AstraZeneca Research & Development 15185 Södertälje, Sweden

Received December 3, 2009

A thorough comparison between different QSAR modeling strategies is presented. The comparison is conducted for local versus global modeling strategies, risk assessment, and computational cost. The strategies are implemented using random forests, support vector machines, and partial least squares. Results are presented for simulated data, as well as for real data, generally indicating that a global modeling strategy is preferred over a local strategy. Furthermore, the results also show that there is an pronounced risk and a comparatively high computational cost when using the local modeling strategies.

1. INTRODUCTION

For the general case of QSAR modeling, the starting point is a data set with chemical structures (compounds represented by a molecular format) and one or more biological activities acquired from experiments. From the compounds, a set of descriptors is calculated generating the descriptor space that is believed to contain properties that can link or explain the differences in the activity.

In the literature, two distinctly different QSAR modeling strategies have been applied, commonly denoted "local" and "global" QSAR models. For example, Guha et al. defines the global model as the model built on the entire training data and that there may be groups of molecules in the data set that has a specific set of features that relates to their activity or inactivity, such a set should in that case represent a local structure−activity relationship.[1] This local set can be extracted by fingerprint or descriptor similarity. Zhang et al.[2] use the Euclidean norm in the descriptor space to determine, which compounds are chemically similar and thereby local. The assumption that molecules that are close in descriptor space or fingerprint space will tend to have the same properties has been studied by Martin et al.[3] They relate fingerprint similarity to biological activity, and this connection seems to be somewhat unclear.

When generating QSAR models on a subset of the available data, examples and thereby information is left out. This raises two important questions. Can one actually gain accuracy by doing this? Are there any risks or drawbacks with this kind of removal of information? In the literature QSAR models based on subsets of the data,[4] as well as all available data,[5] give good accuracy. To the best of the authors knowledge there have not been any work done on QSAR risk assessment. However there has been work estimating errors based on descriptor or compound similarities.[6,7]

This shows that there is a need to test and validate differences in *local* and *global* modeling strategies and how different numerical routines and modeling algorithms can handle the differences. The aim is to gain knowledge about the expected predictive performance of *local* and *global* modeling strategies, to compare different machine-learning algorithms, and to investigate possible risks in terms of the definition and usage of applicability domain of *local* and *global* modeling strategies. It is also interesting to see how the different machine-learning algorithms make use of the available information.

It can be difficult to study behavior of different types of modeling algorithms when using real world data, since the underlying relationship is unknown. To be able to more rigorously test different machine-learning methods and modeling strategies, it is possible to use fictitious data. This technique has been successfully used in other sciences and has been called Twilight-Zone simulations.[8] It is a technique, where a predefined solution to the problem is used. This can be applied to QSAR modeling by drawing descriptors from statistical distributions and deciding on a mathematical function, based on the descriptors, that is to be the response. In this way the exact relationship between the descriptors and the response is known.

The remainder of this paper is organized in method, results, discussion and conclusions. In the method part, the general model building procedure is described together with the definitions of local and global models that are used in this paper. The results section includes a brief description of the different data sets and simulation approaches together with their results. Finally the article ends with Discussion and Conclusions.

2. METHOD

A QSAR model is defined by its response and the descriptors describing the compounds. This sets a limit on the available information. Depending on the information at hand different modeling strategies can be applied. In this work, two general classes of modeling strategies have been applied, denoted *ideal* and *restricted*. For the ideal case, all relevant information to accurately describe the underlying

* To whom correspondence should be addressed. E-mail: ernst.ahlberghelgee@gmail.com.
† Safety Assessment.
‡ Medicinal Chemistry.

10.1021/ci900471e
Published on Web 03/10/2010

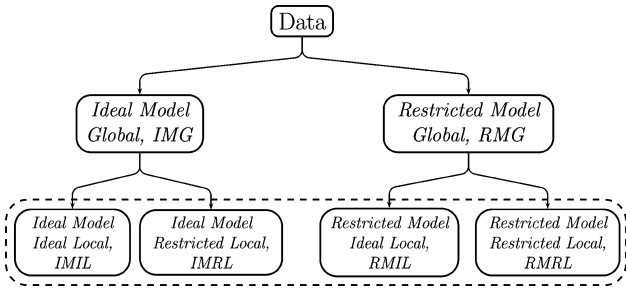


Figure 1. Different modeling strategies that have been applied to the data. The dashed box indicates the local strategies.

relationship is contained by the descriptors, a *complete descriptor set*. For the restricted case, the descriptors are missing relevant information, an *incomplete descriptor set* and cannot be used to describe the underlying relationship but merely an approximation to it. For these two classes either an entire data set, *global*, or a subset of the data, *local*, can be used. This defines a global model as a model built using the entire data set, that is, all available information. A local model, on the other hand, is a model built for a specific example using neighbors from the entire data set. The definition of neighbors can vary, but in this work it is based on a descriptor or fingerprint similarity.

The two global modeling strategies, ideal model global, IMG, and restricted model global, RMG, that have been used are illustrated in Figure 1. IMG uses a complete descriptor set and RMG uses only a subset of these descriptors for model building. For each global modeling strategy, two corresponding local strategies have been applied that locates near neighbors in a restricted or an ideal fashion. Following the structure in Figure 1, in the IMG branch the local strategies both use a complete descriptor set for building models and making predictions. The ideal model ideal local, IMIL, uses the complete descriptor set to identify near neighbors, but the ideal model restricted local, IMRL, only makes use of a restricted subset of descriptors for identifying near neighbors. In the RMG branch, local modeling strategies both use an incomplete descriptor set for model building and predictions, but in the restricted model ideal local, RMIL, they use a complete descriptor set for finding near neighbors and, in the restricted model restricted local, RMRL, use the restricted subset of descriptors for identifying near neighbors.

The global and local modeling strategies above can be summarized as follows: IMG constructs global models, where all the information about the underlying relationship is known and expressed by the descriptors. RMG gives global models that can not correctly describe the underlying relationship. The IMIL case results in local models, where all the information about the underlying relationship is known. IMIL can be directly compared with the RMRL case where the difference is loss of information. RMIL represents the local model case, where external information is added in the neighbor search, which can be relevant in describing the underlying relationship. The addition of this type of information can lead to a model that is really local with respect to the underlying relationship. The IMRL case gives local models, where the local neighborhood is partially unaccounted for or cannot be correctly described, as opposed to RMIL. In fact the underlying relationship is properly described by the descriptors, but the near neighbors have not been selected in accordance to the underlying relationship. The RMRL case results in local models that make use of a neighborhood and a descriptor set that can not completely describe the problem at hand. This can be described as the normal case when building QSAR models since the underlying relationship can not be properly described but one makes use of all information at hand for finding the best possible models and near neighbors.

The different modeling strategies and risk assessments are evaluated using various machine-learning algorithms. To assess individual model performance, a cross-validation approach is used, which is commonly used in literature.[9] A data set is divided into n subsets by a uniform sampling of examples without replacement. Each subset is treated as a test set with the remaining examples as the training set. For each test set, an overall prediction metric is computed. If the response is binary, this metric is defined as the prediction accuracy, and if the response is real valued, the root-mean square error is used instead. The prediction metric is averaged for all test sets.

The generation of global models is straightforward, for each test set a model is built on the remaining examples of the data. Local models are generated for each example in a test set and for each such example near neighbors are retrieved from the remaining examples of the data. Near neighbors are found by using different similarity operators such as Euclidean norm on descriptors or Tanimoto distance on chemical fingerprints. The number of neighbors can either be explicitly set or a cutoff value for the similarity can be used. If a local model can not be built under the specified similarity constraint, the corresponding global model will be used to predict that compound. With the predictions from both local and global modeling strategies at hand, it is possible to directly compare and assess prediction accuracies and errors.

To assess risks of local versus global modeling strategies, the domain of applicability needs to be defined for the local models. By definition a model generated for a test set example, by extraction of near neighbors, is within the *domain of applicability* for that specific example. On the other hand, if such a model is used for any other example, it is used outside its domain of applicability.

3. RESULTS

3.1. Experimental Setup. *3.1.1. Twilight-Zone Simulation.* In the simulation studies, all parameters for the underlying relationship are known, the answer to the problem is known and thereby it is possible to design responses on the basis of different combinations of descriptors and study the effect of local and global modeling strategies more thoroughly.

For this paper descriptors have been drawn from the gamma distribution function resulting in a descriptor set that is believed to mimic the distribution of real chemical descriptors. The simulated descriptor space consists of three different descriptors, d_1, d_2, and d_3, drawn such that $d_1 \in \Gamma(4, 1)$, $d_2 \in \Gamma(9, 1)$, and $d_3 \in \Gamma(7.4, 1)$. The function determining the response is $f_j = \cos(d_{2j} - \bar{d}_2)/(1 + (d_{1j} - \bar{d}_1)^2) + 1.2\sin(1.3(d_{3j} - \bar{d}_3))$, where d_{ij} is the j th point drawn from the i th descriptor above and $\bar{d}_i$ is the mean of the drawn points for that descriptor. On the basis of the above function, regression models and classification models have been built.

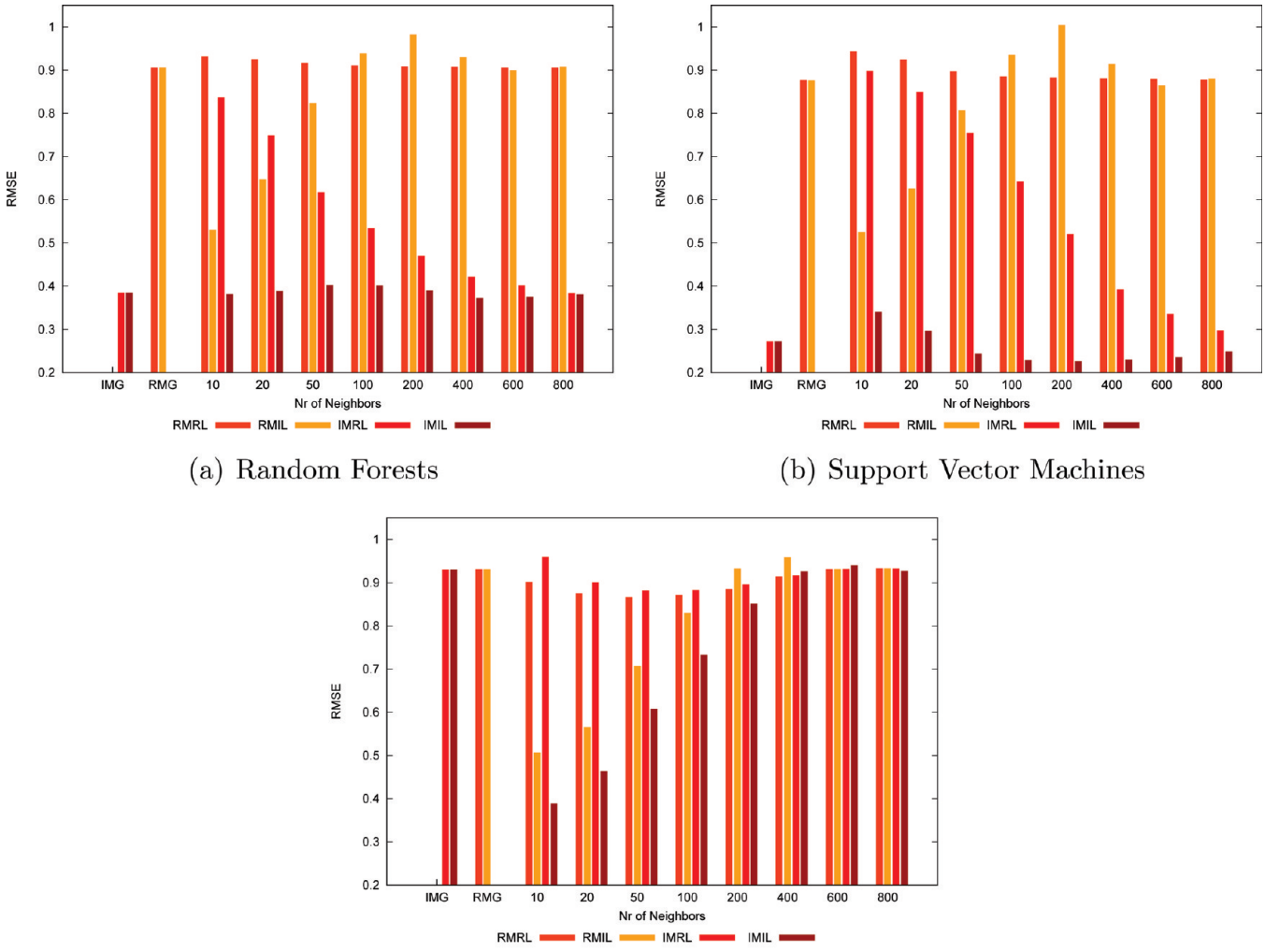


Figure 2. Twilight-zone regression: Root-mean square error, RMSE, of the different machine-learning algorithms for the different local model cases and their respective global counterparts, within the applicability domain.

For classification the response has been changed to $f_{j\text{class}} = I_{f_j>0}$, where I is an indicator function returning one if the statement is true and zero otherwise.

For each modeling strategy 10 seeds have been used and for each seed 1000 examples have been generated. [Seed is a number used to initialize a pseudo random number generator so that it will produce the same sequence of random numbers each time the same seed is used.] The examples have been drawn uniformly into 10 bins and each bin has been used as a test set with the remaining data as a training set. This results in 100 global models for each case and 10000 local models since for each point in the respective test sets a local model has been built. The modeling strategies have been tested using 10, 20, 50, 100, 200, 400, 600, and 800 near neighbors from the training set. The near neighbors have been selected using the Euclidean norm as a distance metric in descriptor space. The results of the simulations are presented as the averaged accuracy or root-mean-square error for each case.

All simulations were conducted in R[10] using the machine-learning libraries e1071[11] for support vector machines, randomForest[12] for random forests (RF), and pls[13] for partial least squares (PLS). Support vector machines, SVM, are very sensitive to parameter optimization and therefore the SVM models have been optimized using a grid search over the γ parameter (2^n, $n = [-5{:}0]$), for both regression and classification and the ε parameter (2^n, $n = [-5{:}-1]$) for regression. The γ parameter is the exponent in the radial basis kernel function and ε is the tolerance of the termination criterion, controlling the width of the loss-insensitive zone in the loss function.

3.1.2. Real World Data. For the real world data the IMG and IMIL cases can not be constructed since the underlying relationship is partially unknown and thus the following strategies are the only that can be applied to the real world data:

(1) The global modeling strategy, RMG, since the underlying relationship of the problem areas are unknown and the descriptors contain information that can only partially describe that relationship.

(2) A pseudo-RMIL modeling strategy, where the neighborhood is defined by Daylight fingerprints. In fact the use of fingerprints could be compared to either RMIL or RMRL depending on if relevant additional information is added by the fingerprints or not.

(3) The RMRL modeling strategy.

In the real world data case, three different data sets were used, Ames[14] mutagenicity data[15] (4253 examples), and

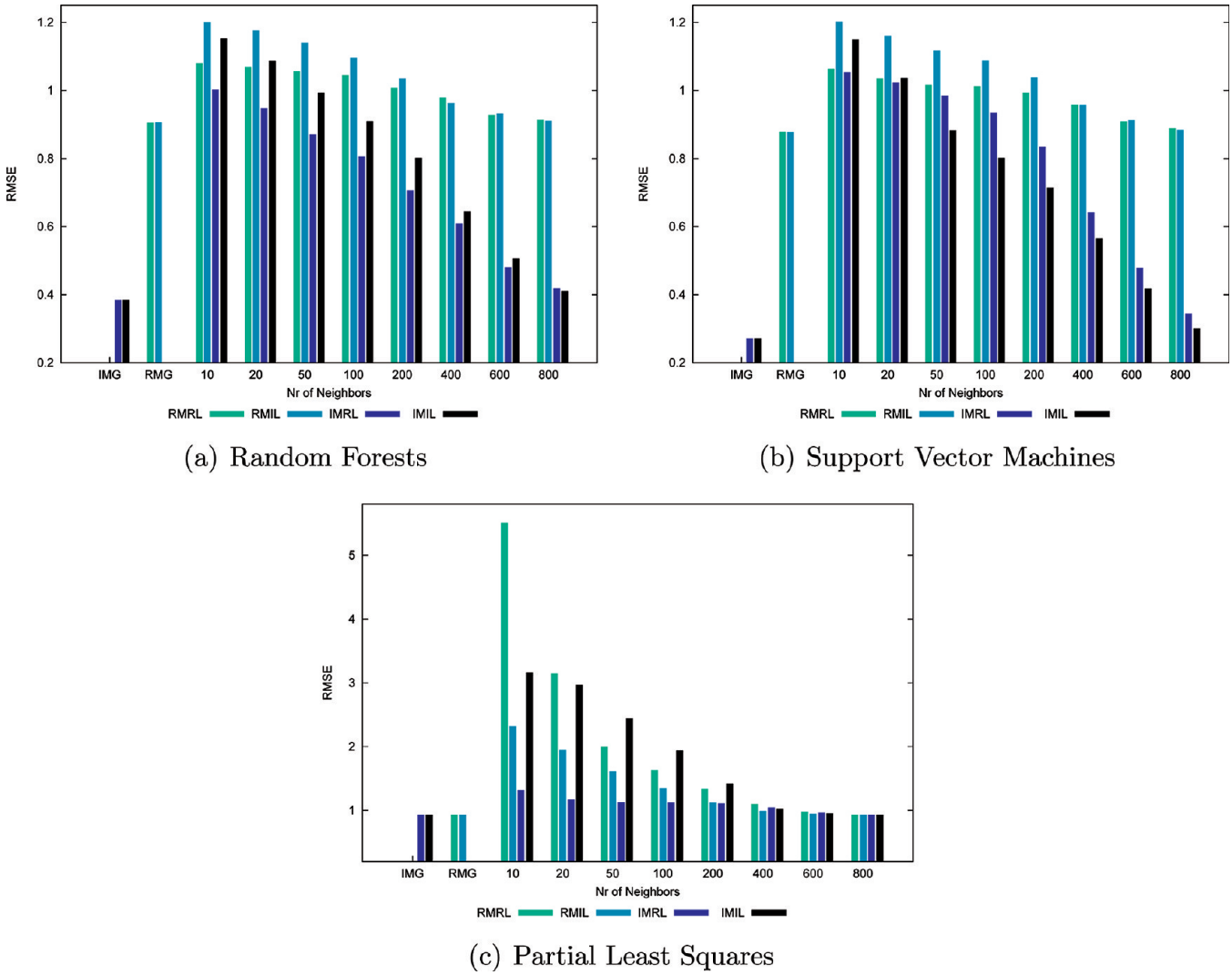


Figure 3. Twilight-zone regression: Root-mean square error, RMSE, of the different machine-learning algorithms for the different local model cases and their respective global counterparts, outside of the applicability domain. NB the scale on the *y*-axis differs for the PLS graph compared to the corresponding RF and SVM graphs.

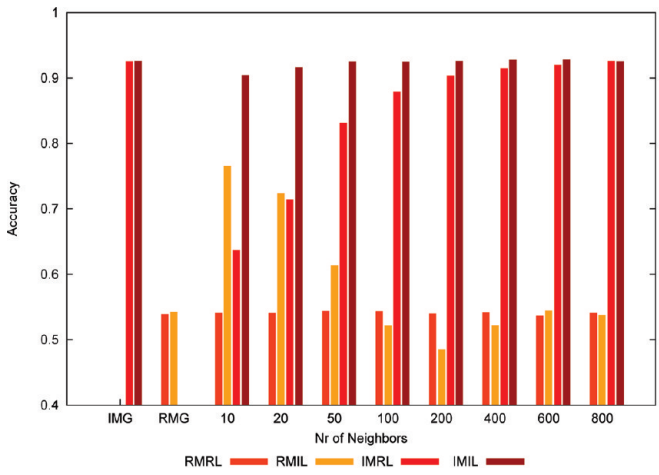
AstraZeneca in-house data sets for hERG, human ether-a-go-go related gene (6020 examples), and solubility data (1186 examples). For all three data sets, binary responses were used. This was accomplished by introducing cutoff values for class definitions. The cut-offs were 30 μ**M** for solubility and a pIC_{50} 5.5 for hERG and positive or negative response for the Ames test, respectively. These three data sets have been used to build local and global classification models. Two sets of descriptors have been used, Signatures[16] and Oe-Selma, an in-house implementation following the work of Labute.[17]

Near neighbors were defined in two different ways: First, using the Euclidean norm as distance metric in descriptor space and second by computing Daylight fingerprints[18] and defining a similarity threshold. In the case where the Euclidean norm defines near neighbors the M nearest neighbors were used as a training set, where M is an integer between ten and the number of examples in the training set. For the fingerprints, a local model was built if the amount of near neighbors with similarity greater or equal to 0.7 was between 10 and 100; otherwise, the prediction of the global model was used.

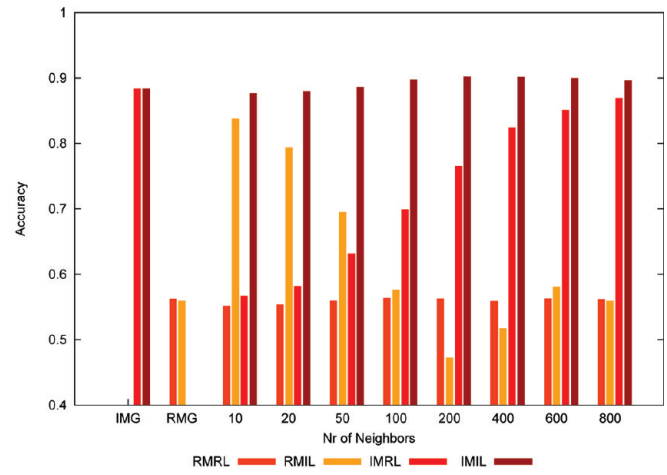
To the real data a set of machine-learning algorithms have been applied. RF, SVM, and PLS models have been computed through an in-house version of Orange,[19] where the SVM relies on libSVM,[20] RF on OpenCV,[21] and PLS on the pls package in R.[13] The parameters used for training the models can be found in the Supporting Information.

3.2. Modeling Strategy Comparisons. The results of the modeling strategy comparisons are presented in three sections, first the simulations using regression models, second the simulations using classification models and third the classification models on real world data. Each modeling strategy is deployed for all simulated data sets and the two local strategies based on RMIL and RMRL are deployed for real data, together with the global modeling strategy RMG.

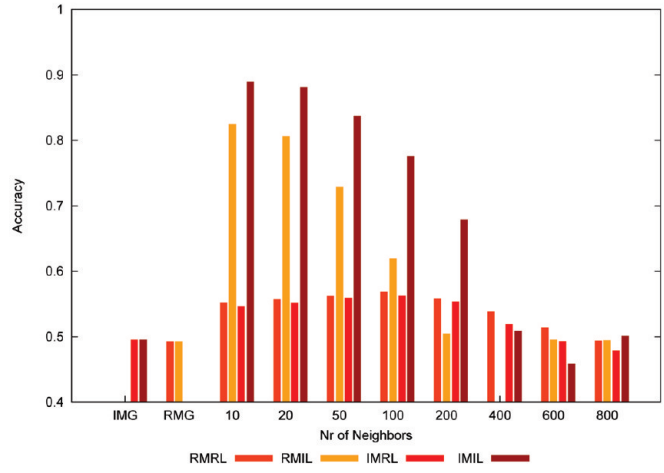
3.2.1. Twilight-Zone Regression Models. The results from the regression models are presented in a series of histograms. Figures 2 and 3 display the averaged overall root-mean square errors of the global model and the local models using different number near neighbors. Figure 2 displays the performance within the applicability domain, and Figure 3 shows the performance outside of the applicability domain. In Figure 2, the RMRL case gives roughly the same errors



(a) Random Forests



(b) Support Vector Machines



(c) Partial Least Squares

Figure 4. Twilight-zone classification: Predictive accuracy of the different machine-learning algorithms for the different local model cases and their respective global counterparts, within the applicability domain.

as its global counterpart, RMG, but the RMIL shows an increased accuracy for all methods. IMRL shows increased errors for all algorithms when the amount of neighbors is small. The IMIL case shows no predictive gain for RF or SVM. SVM seems to work better with a medium sized data set resulting in a slightly lower RMSE when the number of neighbors is in the range of 100−400. For PLS, however, there is a substantial difference in RMSE using only a small amount of near neighbors.

In Figure 3, there is a significantly increased error for the local models compared to the global models for both RF and SVM and for PLS it is only moderately increased, on average. Note however that the errors for RF and SVM are significantly lower than for PLS.

3.2.2. Twilight-Zone Classification Models. The results from the classification models are presented in the following Figures.

Figures 4 and 5 display the overall accuracy of the global model and the local models using different numbers of near neighbors. Figure 4 displays the performance on the local applicability domain, and Figure 5 displays the performance outside of the local applicability domain. In Figure 4, the RMRL series shows a marginal gain in accuracy for PLS but not for RF and SVM. RMIL shows a gain in accuracy for all methods, whereas IMRL shows loss in accuracy for RF and SVM. Only a marginal gain is shown for the PLS method. IMIL display no accuracy gain for RF and SVM but a huge gain for PLS.

3.2.3. Real World Classification Models. The results of the real world data studies, show the same trends as the results from the Twilight-Zone examples. Figure 6−8 show the accuracies for the different machine-learning models on the different data sets.

For the local models, where neighbors were selected using fingerprints, it was not always possible to train a model and the performance of those on the examples that could be predicted is listed in Table 1, where the corresponding coverage of the models also is presented. In the figures, the series "Local Tanimoto + Global" is used where the global model has been used when a local model could not be built.

Figure 6a and 6b show how the different machine-learning methods perform with increasing number of near neighbors for the ames data set. In this figure, the global model is slightly more accurate for all machine-learning methods. Figure 6c and 6d show the results for the "Local Tanimoto + Global" models, which for RF and SVM perform in parity with the global model.

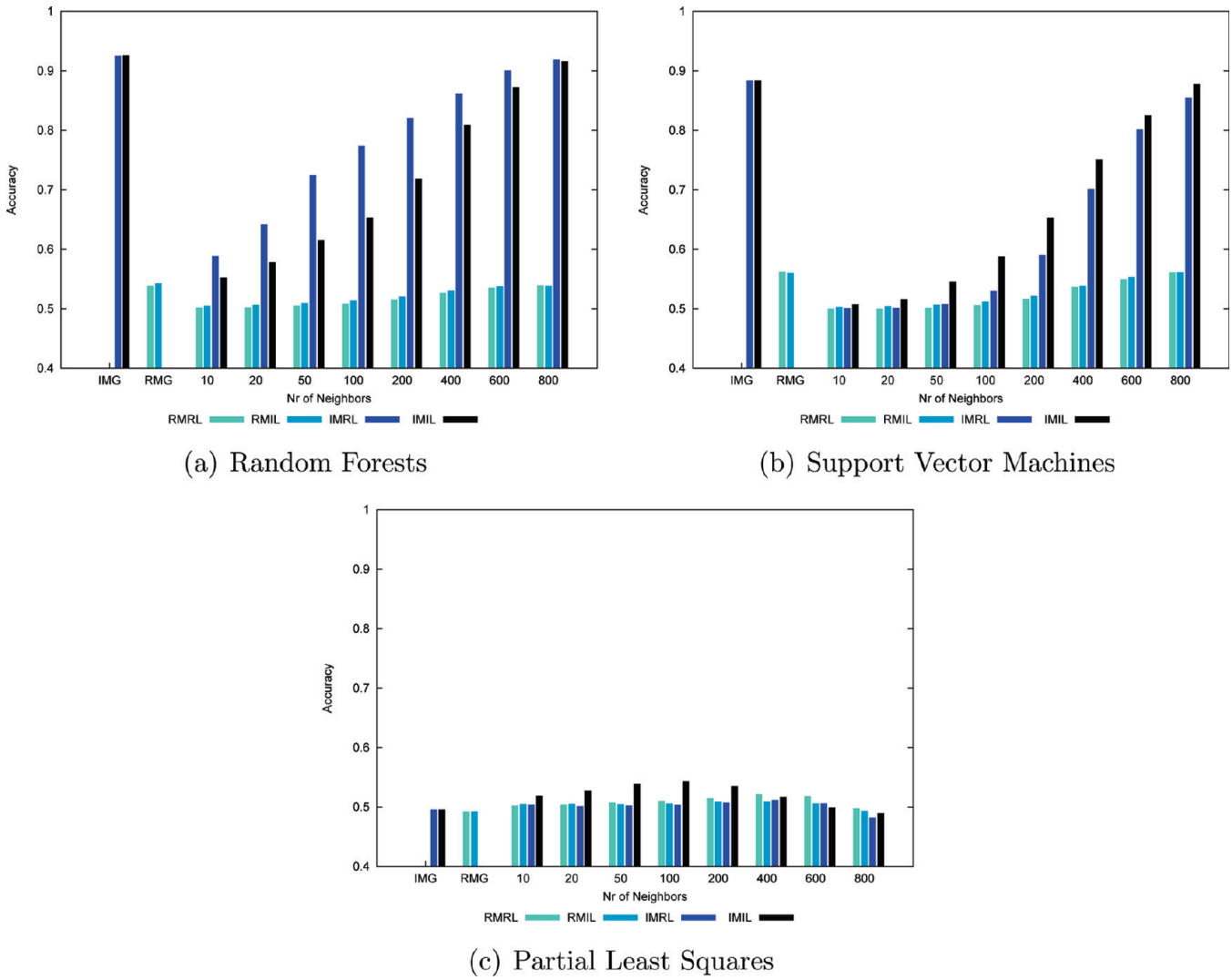


Figure 5. Twilight-zone classification: Predictive accuracy of the different machine-learning algorithms for the different local model cases and their respective global counterparts, outside of the applicability domain.

Figure 7a and b show how the different machine-learning methods perform with increasing number of near neighbors for the hERG data set. All machine-learning methods show about the same performance as for the Ames data set. Figure 7c and 7d show the results for the "Local Tanimoto + Global" models, which for RF and SVM perform in parity with the global model.

Figure 8a and b show how the different machine-learning methods perform with increasing number of near neighbors for the solubility data set. For these plots, the global model is at least slightly more accurate for all machine-learning methods. It is interesting to note that PLS have a large performance drop for the local models with few near neighbors. Figure 8c and d show the results for the "Local Tanimoto + Global" models, which for RF and SVM perform in parity with the global model.

3.3. Modeling Strategy Risk Assessment. *3.3.1. Twilight-Zone Regression Models.* Figure 9 displays the percentage of the cases where the local model is more accurate than the global model in predicting within and outside of the applicability domain of the local model for each machine-learning method respectively. Looking at the trends in the data reveals that outside of the applicability domain the local models perform better as the number of neighbors increases. Within the local applicability domain, the RMRL is not affected by the number of neighbors, it is more accurate than the global model in roughly 50% of the cases. The RMIL is also poor when predicting outside of the local applicability domain. In the IMRL series, it is interesting to note that RF and SVM perform better as the number of neighbors increases. However, for PLS, the IMRL case has a similar performance as the RMRL case. The IMIL models show no gain compared to the IMG model for the RF method, and for the SVM, it can be seen that it is less accurate with few neighbors and performs better with many neighbors, the theory behind those differences is however beyond the scope of this work. Figure 2c clearly shows that for PLS there is a distinct gain for the RMIL and IMIL series compared to the IMG model. This is because the IMG series cannot utilize the extra information and performs approximately as its RMG counterpart.

3.3.2. Twilight-Zone Classification Models. Figure 10 displays the percentage of the cases where the local model is more accurate than the global model when predicting within and outside of the applicability domain of the local models, and Figure 11 shows the opposite, that is, when the global model is more accurate than the local model. This distinction is needed for the classification case; since it is

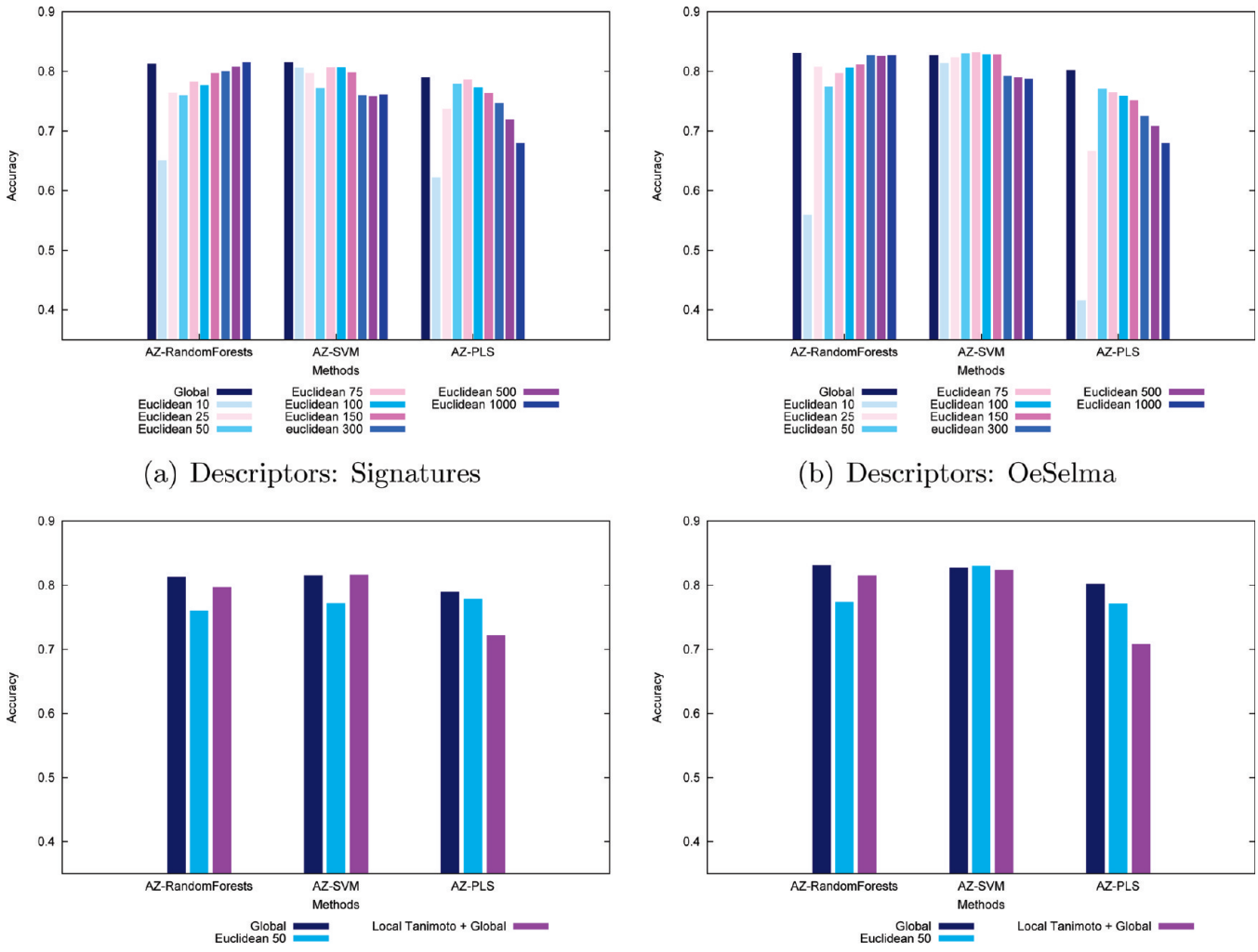


Figure 6. Performance of the different machine-learning algorithms applied to the Ames data set.

likely that the global and local models will give the same prediction for a specific example, this number is left out from the current presentation. This number together with the presented numbers add up to 100%. In general, Figure 10 shows that the higher the number of neighbors the better the local model predicts outside of the local applicability domain, but at the same time, the global character of the model increases and the local gain in predictivity disappears. It is also interesting to note that the local models only perform better than their global counterpart in less than 40% of the cases. The only test case that has a better percentage within the local applicability domain is RMIL. Outside of the applicability domain the global models are much more predictive than the local models, and in roughly 30% of the cases, the global models are more accurate than the local models within the applicability domain as seen in Figure 11. For comparison, the local models are more accurate than their global counterpart within the applicability domain in only 7% of the cases for RF and SVM and about 25% of the cases for PLS.

Figure 5 indicates that for RF and SVM all local models are less accurate compared to their global counterpart, which shows the risk of using local models. For PLS, there is not much of a performance difference between the global and local models; it is however interesting to note the differences between the IMG and RMG models for the different methods.

3.4. Computational Costs. The computational effort for neighbor extraction, building and predicting the local models is shown in Figure 12 for both regression and classification models. The figure indicates that there is a substantial growth in CPU time needed as the number of neighbors increases, which indicates that building local models is time-consuming. Here it is important to remember that a local model is built for each query, thus building local models using 800 near neighbors almost amounts to building as many global models as there are queries. Figure 13 shows the building time in seconds for the models, if the time needed for finding neighbors is added the costs of the local models are approximately doubled. From the figure, it can be seen that training local models using 100 near neighbors results in a 10-fold increase, compared to the global models, in the use of computational resources.

4. DISCUSSION

For a predictive modeling system, there is an interest in being able to predict all incoming compounds. When doing predictive modeling, the model with the highest overall

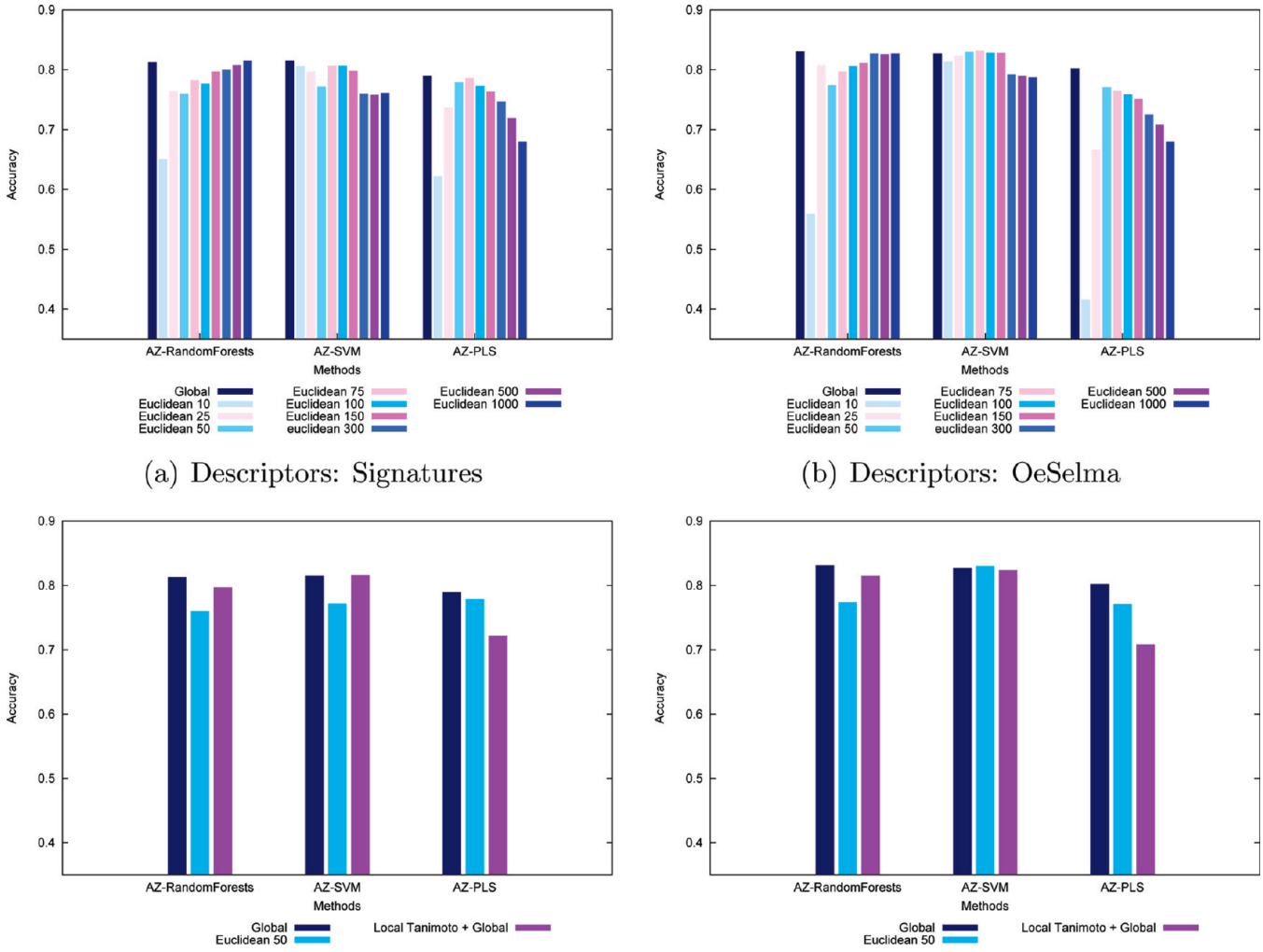


Figure 7. Performance of the different machine-learning algorithms applied to the hErg data set.

accuracy is most commonly the best and preferred model. Sometimes this approach does not lead to an accurate enough model, and in an attempt to overcome this problem, models based on a subset of the data are built.[22] The subset should then capture the problem in a more accurate way. This paper questions the use of subset models for predictive modeling on three major points: (1) There is no statistically validated improvement in accuracy for local models. (2) The risk of falling outside of the applicability domain of the local model is high. Additionally, outside the applicability domain the accuracy of the local model is very poor compared to the accuracy of the global model. (3) In this study, there appears to be a substantial increase in computational cost associated with the local models.

The obvious questions are as follows: How can the modeler be sure that future compounds will fall within the applicability domain and that this domain really is relevant for the issue of interest? If several subset models are built, which one is to be trusted?

4.1. Modeling Strategy Comparisons. The results show that a local modeling strategy only is better than a global strategy if additional information, which is relevant for the underlying relationship, is added in the neighbor search. If a local model, according to the definition used here, performs better than a global model it is advisible to add that additional information to the global model and by doing that the global model can be improved, and there will no longer be a gain from using the local model. This can be seen in the results from the simulated data. Figure 4a compares RMG and RMIL where relevant information has been added in the neighbor search for the local model, and it shows an increased accuracy. The global model updated with the same information is IMG, which has a substantially higher accuracy than RMIL.

Figures 2 and 4 show a substantial difference in RMSE for the different PLS models which is because the PLS IMG cannot utilize the extra information and performs approximately as its RMG counterpart, which indicates that the additional information in the ideal compared to the restricted case is of nonlinear nature with respect to the response.

4.1.1. Twilight-Zone Simulation. Figure 2 shows the performance of the local models on the local test set for the regression models, and Figure 4 shows the performance of classification models on the same data. These figures show the same trends; however, the performance measure for regression is RMSE and its counterpart for classification is accuracy. By analyzing these figures, it is possible to see how the different cases perform. The IMIL series describes what happens when all information is used. This case is more

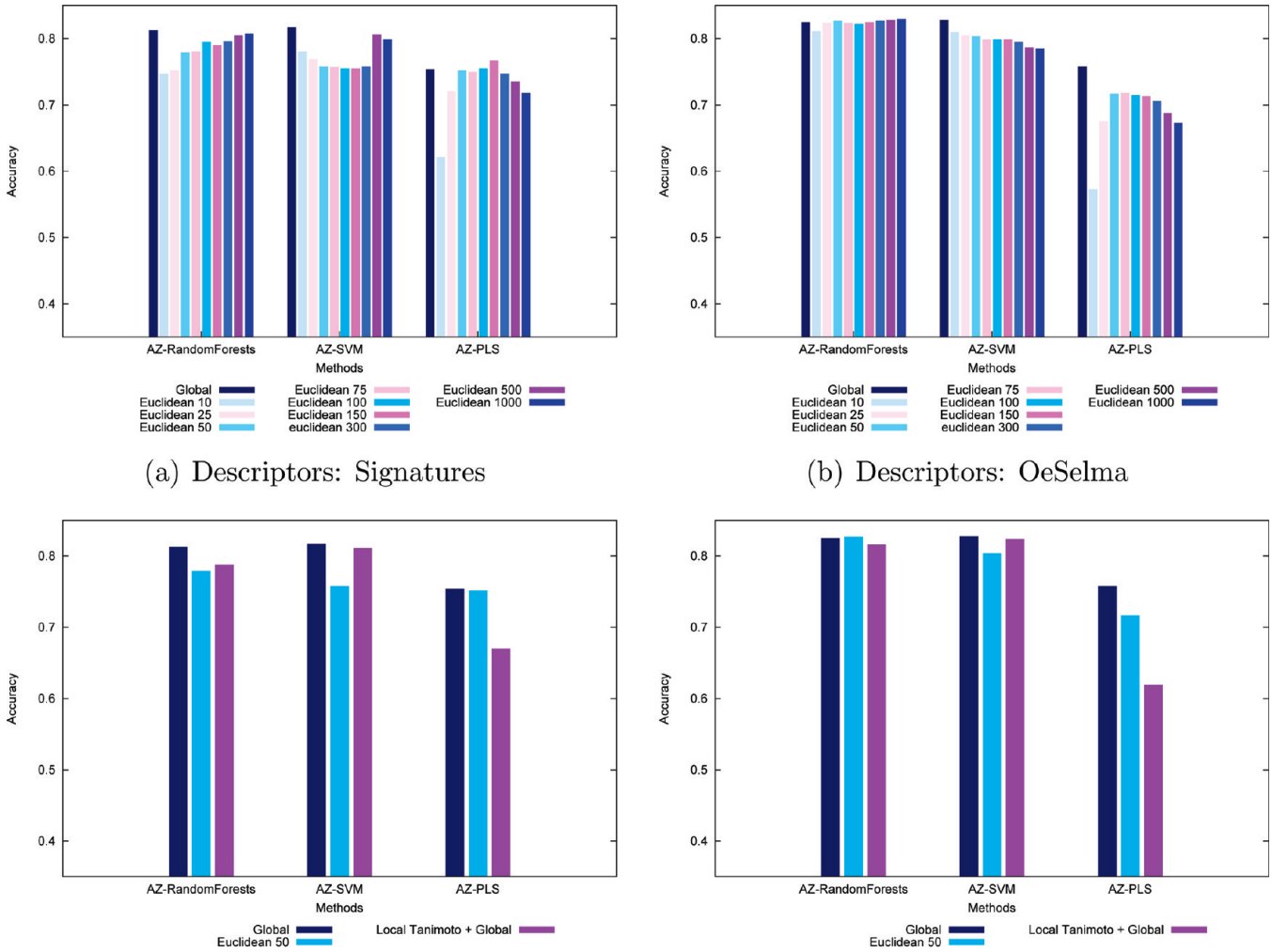


Figure 8. Performance of the different machine-learning algorithms applied to the solubility data set.

Table 1. Accuracy of the Local Tanimoto Models for the Examples Where a Local Model Could Be Built and the Corresponding Coverage

Ames	signatures local Tanimoto	signatures coverage	OeSelma local Tanimoto	OeSelma coverage
az-RandomForests	83	78	84	78
AZ-SVM	91	78	89	78
AZ-PLS	58	78	45	78
hERG	**signatures local Tanimoto**	**signatures coverage**	**OeSelma local Tanimoto**	**OeSelma coverage**
AZ-RandomForests	78	52	82	52
AZ-SVM	82	52	83	52
AZ-PLS	6	52	49	52
solubility	**signatures local Tanimoto**	**signatures coverage**	**OeSelma local Tanimoto**	**OeSelma coverage**
AZ-RandomForests	77	46	78	46
AZ-SVM	82	46	8	46
AZ-PLS	6	46	44	46

of a philosophical question because with all information at hand there is no need for a QSAR model, since the knowledge of the underlying relationship is required to assess the completeness of the information at hand.

The IMRL series describes the case where the neighbors are extracted from a relationship that can not properly describe what is local in the response domain. The result of this is that an external error is introduced and the local models end up less accurate than the global model. This stresses the importance of knowing what is local in the response domain, which in itself is a piece of information that the modeler does not have, unless the relationship is already known and indicates that retrieval of local neighbors based on an incomplete descriptor set can be misleading. The third series is the RMIL series that describes the situation where external information, which is important for the underlying relationship, is added when deciding on what is local and that important external information adds predictivity and value to the model. In this case there is a gain in trying the local model, however if the external information can be added to the model, it is shown from case IMG that the global model is more accurate in predictivity than the local model. This case shows that it is important to describe the problem in different ways, for example with different descriptors, but the global model should always be updated with the most predictive information, generating a more accurate model.

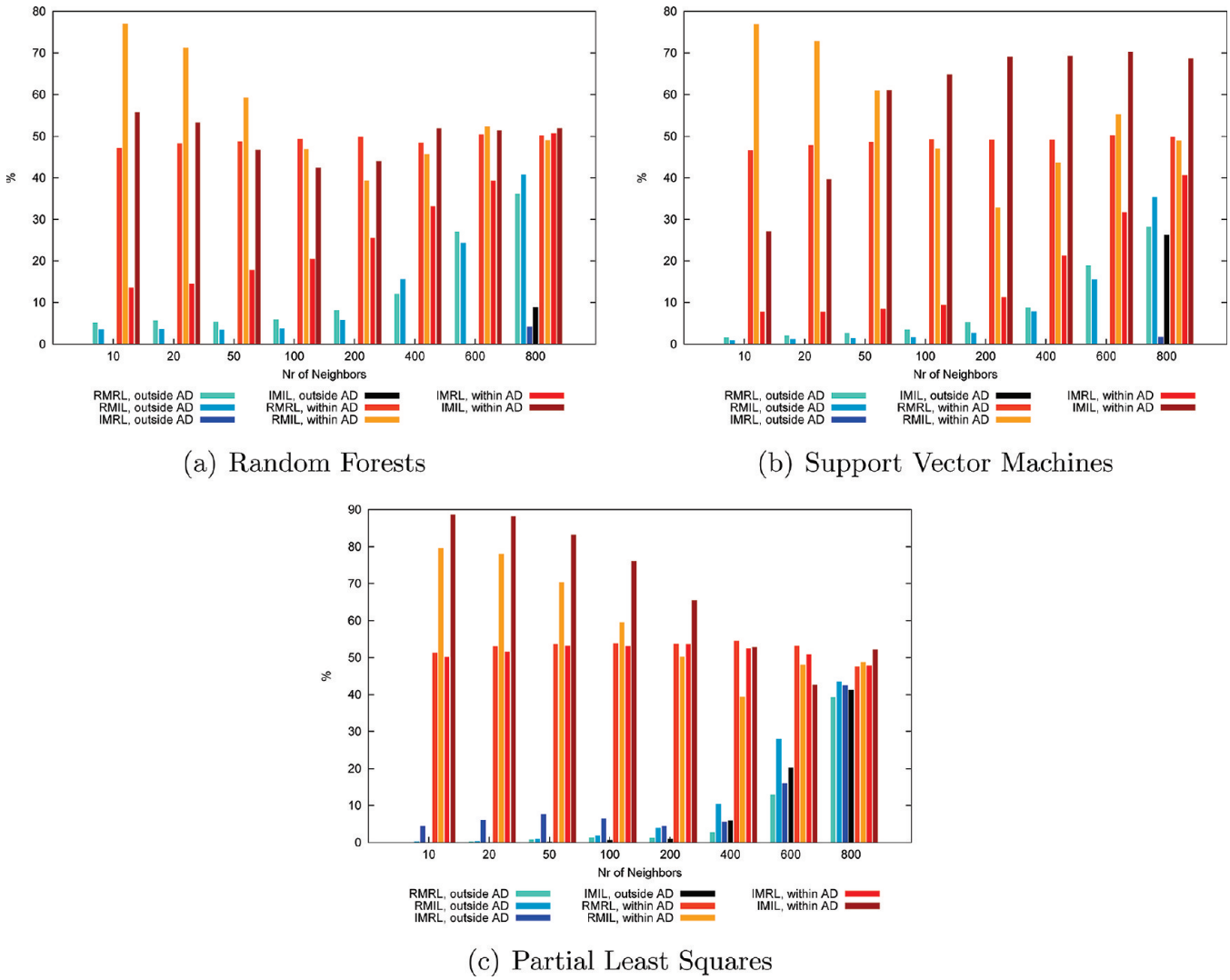


Figure 9. Twilight-zone regression: The percentage of cases where the local model is more accurate than the global one for predictions within and outside of the applicability domain, AD, of the local models.

The fourth series describes the RMRL case, the case where no information is added. This case represents the real world case, since the full relationship is not known when the modeler intends to build a QSAR model, instead all available information is used and then for local models the near neighbors are extracted based on the knowledge and information known to the modeler, which unfortunately does not add information to the system. This implies that the modeler should always choose to build a global model using all information available at the time.

4.1.2. Real World Data. In Figures 6−8, it is interesting to note that the most predictive global model performs better than the local models in the studies for each data set. There are some cases where the models built on neighbors are more accurate when 500−1000 neighbors are used, although in these cases the model can not be assumed to be particularly local any longer.

4.2. Modeling Strategy Risk Assessment. The predictive gain, as shown above, within the local applicability domain is in many cases very small. With that knowledge it is highly relevant to study the risks of falling outside of the local applicability domain and the effects of extracting near neighbors in a suboptimal way.

Figures 9 and 10 are associated with the risk of using local instead of global models. For the classification case, Figure 10 shows that regardless of the method used there is less than a 30% chance of generating a local model that predicts better in the local domain. There are exceptions to this general rule in the RMIL case, which shows that when the neighborhood is correctly computed according to the underlying relationship and the model is restricted, as defined in the method part, some gain in predictive accuracy is established. In the RMRL and the IMRL cases, the performance of the local models increases with increasing number of neighbors. Figure 10 also shows that the more neighbors that are used the better is the model at predicting outside of the local domain but at the same time it becomes more global and loses its additional predictive power in the local domain. On the basis of the comparison between Figure 10 and 11, it can be questioned if local models should be applied to classification data. For the regression case, Figure 9 shows that the ideal local regression models with few near neighbors perform better than their global counterpart on more than 50% of the cases, but this number drops as the number of neighbors increases. The IMRL series for SVM and RF, in Figure 9, also indicates that it is important that predictions

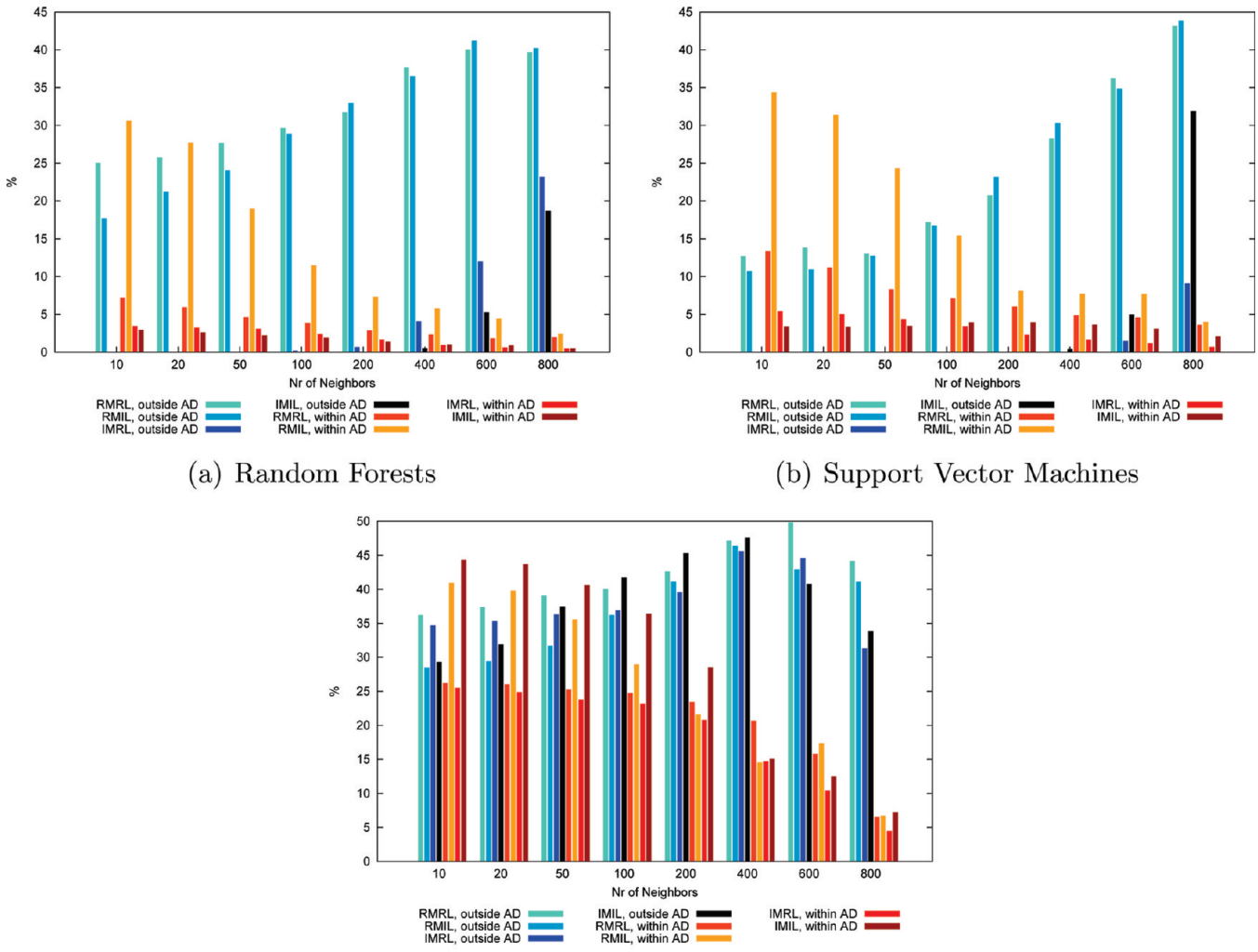


Figure 10. Twilight-zone classification: The percentage of cases where the local model is more accurate than the global one for predictions within and outside of the applicability domain, AD, of the local models.

are made within the domain of applicability. For PLS, however, IMRL has a similar performance as the RMRL case, which shows that PLS is unable to handle the additional information.

The series IMRL shows a case where the local neighborhood is known not to be local with respect to the underlying relationship. Here the local modeling strategy will give less accurate models compared to the global modeling strategy. The RMRL and IMIL cases show that there is no significant difference between the two strategies. Together these results show that applying a local modeling strategy is associated with the risk of using neighbors that are not local, for which prediction performance would degrade compared to the global modeling strategy.

The investigation presented here indicates that using local models for knowledge extraction comes with a high risk. When all the information is used for neighbor extraction, there is a better chance in obtaining a proper local model, but again if the global model has access to the same information, the local model is less accurate, see Figure 2. The results of this investigation indicate that in general, global models should be used for predictive purposes. If, however, a local model is to be used, it is very important that the modeler knows that the compound of interest falls within the local applicability domain, which in many cases cannot be definitively described, and that the local neighbors are chosen in such a way that important information is added and that information needs to be unknown to the model. For example, selecting near neighbors based on the descriptor set used for the model will not add information, but selecting neighbors based on a different set of descriptors, that is able to describe the underlying relationship in a different way add information. Such information, unknown to the model, is generally not available since if it where, the global model could be updated, and there would be no need for the local model.

4.3. Computational Costs. Figure 12 and 13 show that the computational cost for predicting data sets using local models, as they have been defined in this study, is generally higher compared to global models. Thus it appears that to maintain a prediction system using local models a large computational resource needs to be dedicated for these computations.

4.4. Machine-Learning Methods. This study shows that machine-learning methods can be useful for quantitatively describing structure activity relationships. Some algorithms require parameter optimizations to become predictive for example SVM. Of the three methods used, RF seem to be

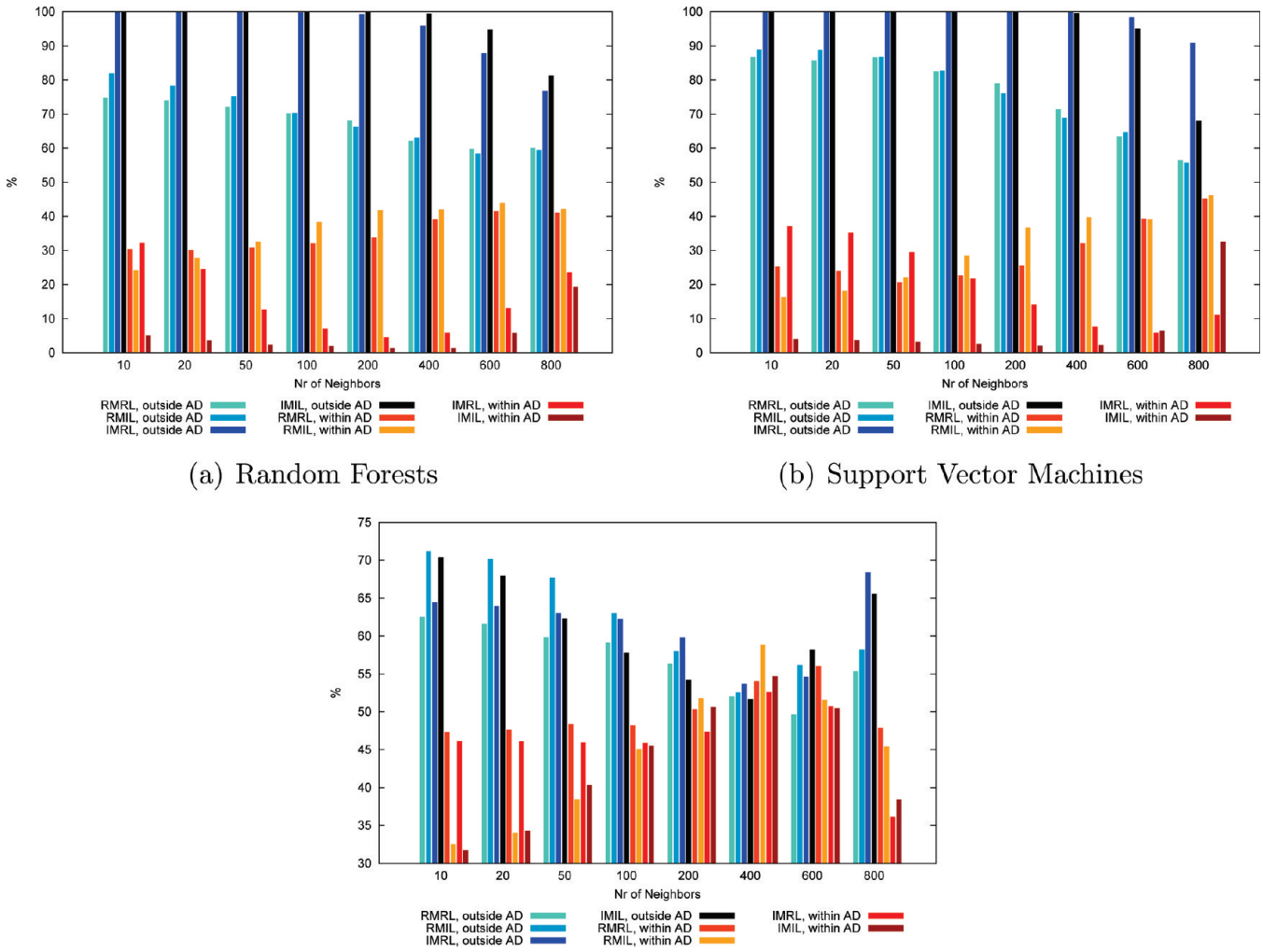


Figure 11. Twilight-zone classification: The percentage of cases where the global model is better than the local one for predictions within and outside of the applicability domain, AD, of the local models.

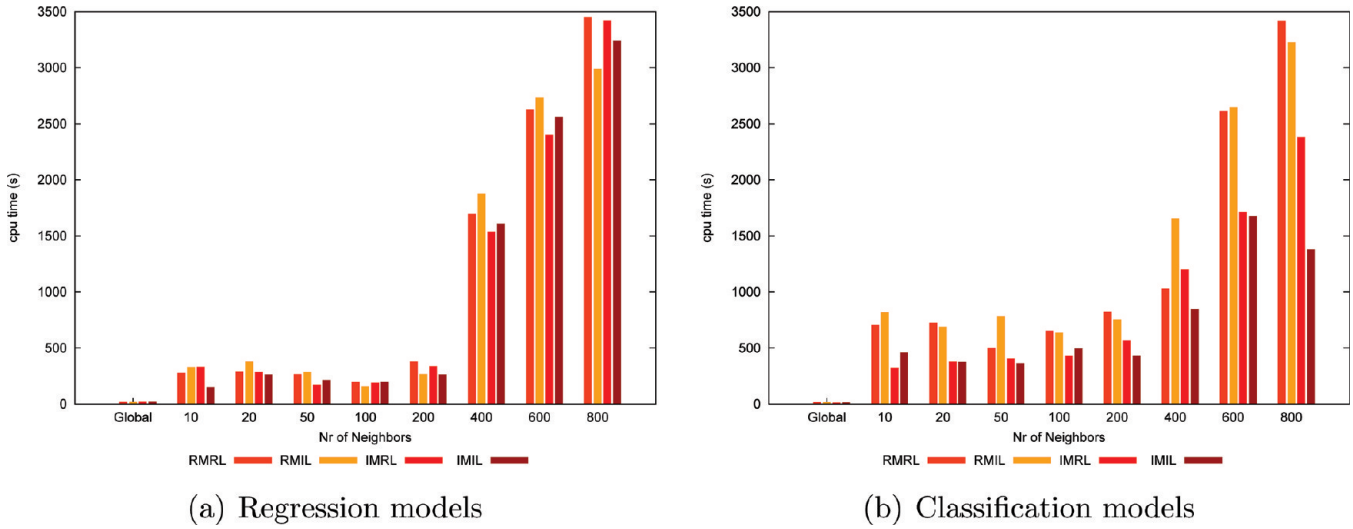


Figure 12. CPU time needed to train the local models with respect to the number of neighbors. The computations where run on a heterogeneous grid.

the most stable algorithm when it comes to delivering accurate models. The results obtained by the PLS algorithm however show larger errors for some cases, especially the simulated data and never had the most predictive model for any data set throughout this study. One reason for this could be that the relationships between the responses and the descriptors used are not entirely linear and since PLS is a linear modeling algorithm that information is not recognized.

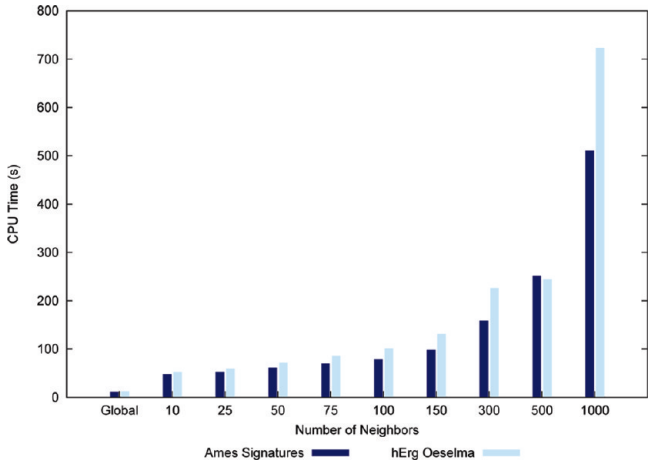


Figure 13. Computational cost for training models.

The fact that PLS is unable to handle nonlinear interactions without first preprocessing the descriptors for higher-order interactions could constitute a drawback for model building.

In a case where the underlying relationship between response and descriptors is linear, SVM, RF, and PLS all exhibit comparable performance, for further details see Supporting Information. However, upon extending the relationship to a nonlinear one, linear methods can not capture the nonlinearity as can be seen by the performance loss for PLS compared to RF and SVM.

The nonlinear machine-learning algorithms perform much better as seen in this work, Figure 2 and 4, and in ref 23. Despite this, linear methods are still widely used. Possible explanations for this are that there has been a lack of easily accessible tools for modeling using nonlinear machine-learning methods, additionally previous limitations in hardware have not enabled handling of large data sets in a global fashion and interpretation of nonlinear models has been difficult. These limitations have traditionally forced researchers to deal with data in subsets where the performance difference between linear and nonlinear machine-learning algorithms has perhaps not been very pronounced.

4.5. Common Use of "Local". It should be noted that the common use of "Local" for models that are concentrated around a series of compounds with the same core structure for which descriptors are computed which cannot extend to the complete data set, for example substituent parameters. This creates an information poor model which cannot make use of the remainder of the data nor can it be used to add information to an existing global model and thereby making it difficult to transfer information from substituent parameters to a global model.

5. CONCLUSIONS

The results from this study, both for the simulated and the real world data, indicate that building local models is associated with relatively high computational costs, high risks in defining local, and the local models give no reliable increased predictive performance. The use of simulated data gives statistical rigor to the comparison of modeling strategies, followed by real data showing similar trends. The results of this study suggests that building global models and keeping them updated with new information that affects the underlying relationship is the best way to consistently ensure accurate models.

Supporting Information Available: Machine-learning parameters used when training models on the real world data, results of two additional test cases for the Twilight-zone simulations, and additional test cases for comparison, where the first test case uses a linear relationship and the second case uses a monotonic nonlinear function. This information is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Guha, R.; Dutta, D.; Jurs, P. C.; Chen, T. Local Lazy Regression: Making Use of the Neighborhood to Improve QSAR Predictions. *J. Chem. Inf. Model.* **2006**, *46*, 1836–1847.
(2) Zhang, S.; Golbraikh, A.; Oloff, S.; Kohn, H.; Tropsha, A. A Novel Automated Lazy Learning QSAR (ALL-QSAR) Approach: Method Development, Applications, and Virtual Screening of Chemical Databases Using Validated ALL-QSAR Models. *J. Chem. Inf. Model.* **2006**, *46*, 1984–1995.
(3) Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do Structurally Similar Molecules Have Similar Biological Activity? *J. Med. Chem.* **2002**, *45*, 4350–4358.
(4) Yuan, H.; Wang, Y.; Cheng, Y. Local and Global Quantitative Structure Activity Relationship Modeling and Prediction for the Baseline Toxicity. *J. Chem. Inf. Model.* **2007**, *47*, 159–169.
(5) Gavaghan, C. L.; Hasselgren, C.; Blomberg, N.; Gert, S.; Boyer, S. Development, interpretation and temporal evaluation of a global QSAR of hERG electrophysiology screening data. *J. Comput.-Aided Mol. Des.* **2006**, *21*, 189–206.
(6) Schultz, T. W.; Hewitt, M.; Netzeva, T. I.; Cronin, M. T. D. Assessing Applicability Domains of Toxicological QSARs: Definition, Confidence in Predicted Values, and the Role of Mechanisms of Action. *QSAR Comb. Sci.* **2007**, *26*, 238–254.
(7) Tetko, I. V.; Bruneau, P.; Mewes, H.-W.; Rohrer, D. C.; Poda, G. I. Can We Estimate the Accuracy of ADMET Predictions? *Drug Discovery Today* **2006**, *11*, 700–707.
(8) Henshaw, W. D. A Fourth-Order Accurate Method for the Incompressible Navier−Stokes Equations on Overlapping Grids. *J. Comput. Phys.* **1994**, *113*, 13–25.
(9) Wold, S. Validation of QSAR's. *Quant. Struct.−Act. Relat.* **1991**, *10*, 191–193.
(10) R Development Core Team *R: A Language and Environment for Statistical Computing*, version 2.7.2; R Foundation for Statistical Computing: Vienna, Austria, 2008 (ISBN 3-900051-07-0).
(11) Dimitriadou, E.; Hornik, K.; Leisch, F.; Meyer, D.; Weingessel, A. *e1071: Misc Functions of the Department of Statistics (e1071)*; TU Wien; 2006, R package version 1.5-16.
(12) Liaw, A.; Wiener, M. Classification and Regression by randomForest. *R News* **2002**, *2*, 18–22.
(13) Wehrens, R.; Mevik, B.-H. *pls: Partial Least Squares Regression (PLSR) and Principal Component Regression (PCR)*; 2007, R package version 2.0-1.
(14) Ames, B.; Lee, F.; Durston, W. An improved bacterial test system for the detection and classification of mutagens and carcinogens. *Proc. Natl. Acad. Sci.* **1973**, *70*, 782–786.
(15) Kazius, J.; McGuire, R.; Bursi, R. Derivation and Validation of Toxicophores for Mutagenicity Prediction. *J. Med. Chem.* **2005**, *48*, 312–320.
(16) Faulon, J.-L. Translator. http://www.cs.sandia.gov/ ~jfaulon/QSAR/translator.tar.gz (Accessed June 12, 2008).
(17) Labute, P. A widely applicable set of descriptors. *J. Mol. Graphics Modell.* **2000**, *18*, 464–477.
(18) Fingerprints. http://daylight.com/dayhtml/doc/theory/theory.finger.html (Accessed August 23, 2009).
(19) Demsar, J.; Zupan, B.; Leban, G. *Orange: From Experimental Machine Learning to Interactive Data Mining*; Faculty of Computer and Information Science, University of Ljubljana, 2004.
(20) Chang, C.-C.; Lin, C.-J. LIBSVM: A library for support vector machines, 2001, http://www.csie.ntu.edu.tw/cjlin/libsvm.
(21) OpenCV. http://opencv.willowgarage.com/wiki/ (Accessed May 28, 2009).
(22) Penzotti, J. E.; Landrum, G. A.; Putta, S. Building predictive ADMET models for early decisions in drug discovery. *Curr. Opin. Drug Discovery Dev.* **2004**, *7*, 49–61.
(23) Carlsson, L.; Ahlberg Helgee, E.; Boyer, S. Interpretation of Nonlinear QSAR Models Applied to Ames Mutagenicity Data. *J. Chem. Inf. Model.* **2009**, *49* (11), 2551–2558, doi: 10.1021/ci9002206.

CI900471E