JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

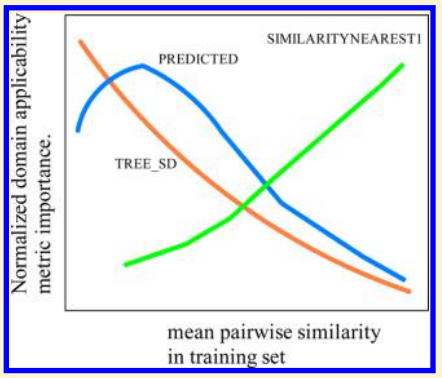
The Relative Importance of Domain Applicability Metrics for Estimating Prediction Errors in QSAR Varies with Training Set Diversity

Robert P. Sheridan*

Cheminformatics Department, RY800B-305, Merck Research Laboratories, Rahway, New Jersey 07065, United States

Supporting Information

ABSTRACT: In QSAR, a statistical model is generated from a training set of molecules (represented by chemical descriptors) and their biological activities (an "activity model"). The aim of the field of domain applicability (DA) is to estimate the uncertainty of prediction of a specific molecule on a specific activity model. A number of DA metrics have been proposed in the literature for this purpose. A quantitative model of the prediction uncertainty (an "error model") can be built using one or more of these metrics. A previous publication from our laboratory (Sheridan, R. P. *J. Chem. Inf. Model.* **2013**, *53*, 2837−2850) suggested that QSAR methods such as random forest could be used to build error models by fitting unsigned prediction errors against DA metrics. The QSAR paradigm contains two useful techniques: descriptor importance can determine which DA metrics are most useful, and cross-validation can be used to tell which subset of DA metrics is sufficient to estimate the unsigned errors. Previously we studied 10 large, diverse data sets and seven DA metrics. For those data sets for which it is possible to build a significant error model from those seven metrics, only two metrics were sufficient to account for almost all of the information in the error model. These were TREE_SD (the variation of prediction among random forest trees) and PREDICTED (the predicted activity itself). In this paper we show that when data sets are less diverse, as for example in QSAR models of molecules in a single chemical series, these two DA metrics become less important in explaining prediction error, and the DA metric SIMILARITYNEAREST1 (the similarity of the molecule being predicted to the closest training set compound) becomes more important. Our recommendation is that when the mean pairwise similarity (measured with the Carhart AP descriptor and the Dice similarity index) within a QSAR training set is less than 0.5, one can use only TREE_SD, PREDICTED to form the error model, but otherwise one should use TREE_SD, PREDICTED, SIMILARITYNEAREST1.



INTRODUCTION

In quantitative structure−activity relationship (QSAR) studies, a statistical model is generated from a training set of molecules (represented by chemical descriptors) and their biological activities. This model is used to predict the activities of new compounds. We call this the "activity model". The subfield of QSAR studies called domain applicability (DA)[1−34] attempts to define the reliability of prediction for a new molecule *M* on a given activity model. One can approach domain applicability in two ways. One approach is to define the chemical space for which predictions are reliable for the activity model. If *M* is within the space, it is reliably predicted; otherwise, it is not. Another approach, which we address here, is to estimate an error bar on the prediction of *M* on the activity model, i.e. a "prediction uncertainty". A larger error bar indicates a less reliable prediction.

A number of metrics have been demonstrated in the literature to be at least somewhat correlated with the reliability of prediction. Such "DA metrics" come in a number of types: "distance/similarity to model", "bagged variance", "local error", etc. One can imaging building another quantitative model, which we can call an "error model", to estimate the prediction uncertainty from one or more DA metrics. It should be noted that some methods of calculating prediction uncertainty do not use explicit DA metrics. For example, refs 4 and 30 use the original chemical descriptors instead of DA metrics. Also, some QSAR methods, such as Gaussian processes,[35−37] intrinsically produce a prediction uncertainty along with the prediction and do not require a separate error model.

In an earlier paper,[24] we showed that QSAR methods themselves, in our case random forest (RF), would be useful in making an error model. Where normally one would use "activity" in QSAR, one would use "prediction error", and instead of "chemical descriptors", one would use "DA metrics". Applying the QSAR paradigm to the construction of an error model has useful features. Descriptor importance can be used to decide which DA metrics are most useful for a given error model. Cross-validation can be used to decide whether a self-consistent error model can be built and which subset of DA metrics is sufficient to explain the errors.

Received: February 27, 2015
Published: May 21, 2015

DOI: 10.1021/acs.jcim.5b00110
J. Chem. Inf. Model. 2015, 55, 1098−1107

In the earlier paper, we examined 10 large, diverse QSAR data sets and built activity models using random forest. When it was possible to build a significant error model, only two DA metrics of the seven we examined were sufficient to account for almost all of the error. These metrics were TREE_SD (the variation of prediction among random forest trees) and PREDICTED (the predicted activity itself). It did not seem to matter whether these data sets were on- or off-target.

For the past 10 years, most of the QSAR models published at Merck for in-house use were based on large, diverse off-target data sets, so TREE_SD and PREDICTED alone should be adequate. Recently, however, the trend has been to publish models built from smaller, less diverse, on-target data sets (some based on a single chemical series), and we may need to modify this approach. In this paper, we examine this question by starting with 15 data sets of on-target activities, selecting progressively less diverse subsets from them, and monitoring the importance of seven DA metrics as a function of diversity. For those data sets for which a statistically significant error model can be built, we find that as the diversity of the training set decreases, DA metrics such as TREE_SD and PREDICTED become less important and metrics such as SIMILARITY-NEAREST1 (similarity to the nearest compound in the training set) become more important. Error models for less diverse data sets need all three metrics.

METHODS

Overview: Activity Models versus Error Models. A conventional QSAR model (here called an "activity model") is built using a QSAR method Q from an activity data set T, which consists of a combination of biological activities and chemical descriptors D (section 1 in Figure 1). Building an activity model may also involve finding optimum values for adjustable parameters P.

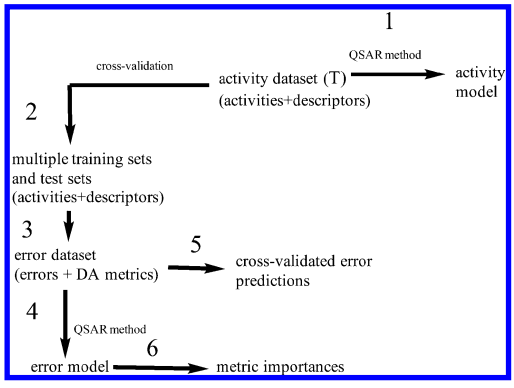


Figure 1. Scheme for activity models and error models. Sections of the scheme are indicated by numbers 1−6.

To build an error model associated with the activity model, one must have a set of predictions on the activity model, with one or more DA metrics associated with each prediction, and the corresponding "errors", which are the absolute differences between the predicted and observed values. We generate the predictions and errors by cross-validation on T using the same D, Q, and P used for the activity model (sections 2 and 3 in Figure 1).

To make an activity prediction for a new molecule M, one scores the chemical descriptors for M on the activity model. To estimate the prediction uncertainty for M, one generates the DA metrics for M using the activity model and/or the chemical descriptors for T and then makes a prediction for M against the error model. The prediction for M can be scaled into a prediction uncertainty.

In this paper we are not so much concerned with producing the prediction uncertainty but rather with which subset of DA metrics is the best to use to build the error model (section 4 in Figure 1) for the appropriate level of diversity in the QSAR training set. The methods in sections 5 and 6 in Figure 1 will help us determine that. Details are provided below.

Data Sets. We will show as examples the 15 in-house QSAR data sets listed in Table 1. Whereas in our previous publication[24] we used a mixture of on- and off-target data sets, all of these are on-target data sets. This choice was made because on-target data sets tend to contain large "series", and as we select less diverse subsets of molecules, large enough sets of similar molecules will be available to make useful QSAR models.

Some of these data sets have a substantial fraction of "qualified data", for example, "$IC_{50} > 30\ \mu M$". Most off-the-shelf QSAR methods, including the implementations of RF used here, do not treat qualified data as anything but a definite value, i.e. >30 μM would be treated as 30 μM, or $-\log(IC_{50}) = 4.5$ in units of molar.

Selection of Subsets of Varying Diversity. Defining diversity requires a definition of similarity. Throughout we define similarity using the Carhart AP descriptor and the Dice similarity index[38] to be consistent with our previous work. However, we expect other other similarity standards based on substructure descriptors, such as ECFP4/Tanimoto,[39] to be equally useful. Different descriptor/index combinations, even though their lower and upper limits will be 0 and 1, have different distributions of values and different cutoffs for what a chemist would consider "similar". For AP/Dice, randomly selected pairs of compounds have a mean similarity of 0.28, and compounds with a similarity of ≥0.65 would be considered clear analogues by most chemists.

A cartoon illustrating the selection of subsets by clustering is displayed in Figure 2. Let T' be the set of all compounds for which we have activity measures. The data sets in Table 1 would be examples. We cluster each T' by the Butina algorithm.[40] This is a sphere exclusion algorithm that uses a similarity cutoff. The molecule with the most neighbors ("neighbors" have similarity greater than or equal to the cutoff) becomes the "centroid" of the first cluster, molecules more similar to that centroid than the cutoff become members of the first cluster. The molecule with the most neighbors not already in the first cluster becomes the centroid of the second cluster, etc. For our purposes, the cluster with the most molecules (usually the first cluster) is taken as training set T for further processing (see below). Since we can use several cutoffs to vary the diversity (i.e., 0.0, 0.3, 0.4, 0.5, 0.6, 0.7), we give names to the resulting subsets T of the form *T0.0*, *T0.3*, etc. A cutoff of 0.0 includes all of the molecules in the largest (and only) cluster, so *T0.0* is the same as T'. As the similarity cutoff rises, T contains fewer molecules that are more similar to each other, i.e., T becomes less diverse. The method of clustering does not matter as long as sets of T of varying diversity can be produced. The actual diversity within T is measured by the mean pairwise similarity (MPS) using the AP descriptor and the Dice similarity index. A T with MPS < 0.35 could be considered "diverse", and a T with MPS ≥ 0.65 could be considered "not diverse".

DOI: 10.1021/acs.jcim.5b00110
J. Chem. Inf. Model. 2015, 55, 1098−1107

Table 1. Data Sets

name	description	*N* in *T0.0*	CV-R^2 activity for *T0.0*	CV-R^2 SQRT(UE) all DA metrics for *T0.0*
5HT2	$-\log(IC_{50})$ for binding to the 5-HT2 receptor	7308	0.74	0.35
ADENOSINE	$-\log(IC_{50})$ for binding to the adenosine-2 receptor	5018	0.43	0.18
AII	$-\log(IC_{50})$ for inhibition of the angiotensin-II receptor	2763	0.78	0.36
BACE	$-\log(IC_{50})$ for inhibition of β-secretase-1	17469	0.75	0.15
CB1	$-\log(IC_{50})$ for binding to the cannabinoid-1 receptor	11637	0.57	0.12
COX2	$-\log(IC_{50})$ for inhibition of cyclooxygenase-2	1178	0.24	0.10
DPP4	$-\log(IC_{50})$ for inhibition of dipeptidylpeptidase-4	6905	0.59	0.23
ERK2	$-\log(IC_{50})$ for inhibition of ERK2 kinase	12843	0.44	0.41
FACTORXIA	$-\log(IC_{50})$ for inhibition of factor XIa	9536	0.67	0.44
HIVINT	$-\log(IC_{50})$ for inhibition of HIV integrase	2413	0.49	0.16
HIVPROT	$-\log(IC_{50})$ for inhibition of HIV protease	4311	0.74	0.22
HIVRT	$-\log(IC_{50})$ for inhibition of HIV reverse transcriptase	9571	0.79	0.24
MBLI	$-\log(IC_{50})$ for inhibition of metallo-β-lactamase	643	0.55	0.18
NK1	$-\log(IC_{50})$ for binding to substance P receptor	13482	0.67	0.15
THROMBIN	$-\log(IC_{50})$ for inhibition of thrombin	6800	0.69	0.16

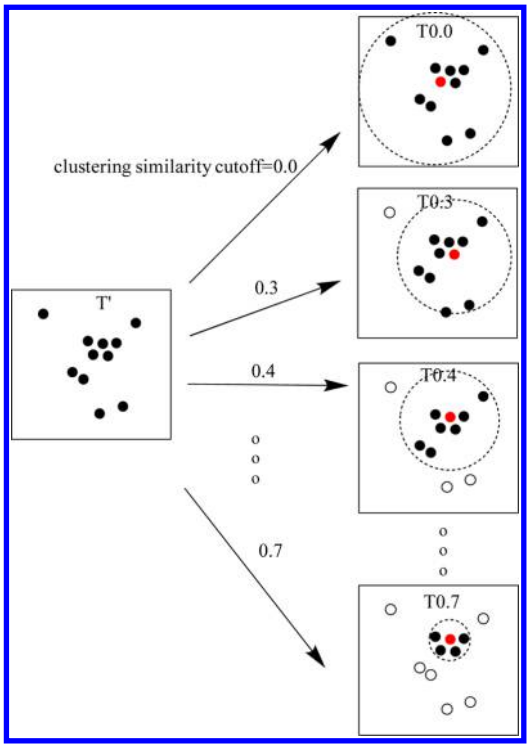


Figure 2. Generating training sets of lower diversity by clustering with higher similarity cutoffs. In this type of diagram representing chemical descriptor space, smaller distances between molecules represent higher similarity. Filled circles are molecules in the largest cluster, with the "centroid" of the cluster in red. The dashed circle represents the similarity radius around the centroid.

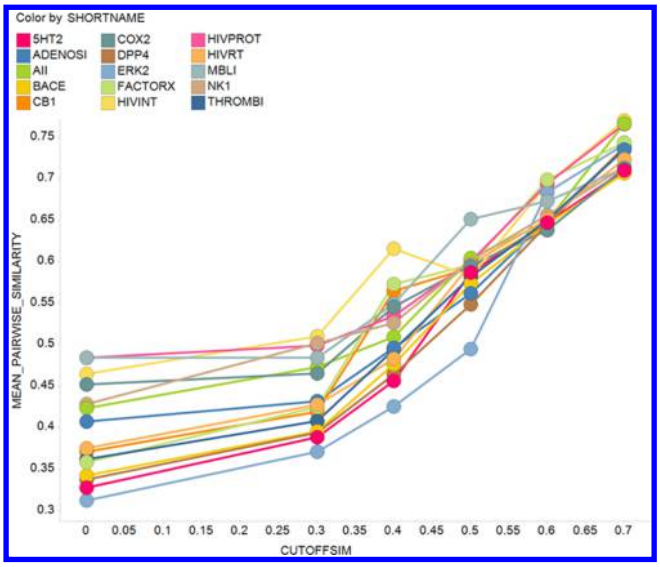


Figure 3. Mean pairwise similarity (MPS) of molecules within subsets *T* as a function of the similarity cutoff for the clustering algorithm. Data sets are distinguished by color. A subset with an MPS of ~0.3 could be considered maximally diverse. A mean pairwise similarity of ≥0.65 would indicate a subset consisting of close analogues.

The MPS as a function of the similarity cutoff is shown in Figure 3. Clearly different data sets can have different MPSs for any given cutoff, but overall there is a trend toward higher MPS as the cutoff rises. For any given data set, the change in MPS may not necessarily be strictly monotonic with higher cutoff (HIVINT is an example). This is the case because within a data set there can be more than one series with almost equal numbers of molecules, and changing the cutoff can cause a different series to be chosen as the largest cluster.

QSAR Methods and Descriptors for the Activity Model (Section 1 in Figure 1). Details of how we generate activity models are provided in our previous publication.[24] Our QSAR method of choice is random forest.[41] RF is very attractive for DA purposes because, as an ensemble method (the model consists of 100 recursive partitioning "trees," each of which produces a separate prediction), we get a "bagged variance" DA metric with no extra work. As in the previous paper, we are using for the activity models the union of the Carhart atom pairs (AP)[38] and a donor−acceptor pair (DP), called "BP" in Kearsley et al.,[42] which we have found to give the most accurate cross-validated predictions for activity models.

Construction of Error Data Sets (Sections 2 and 3 in Figure 1). In order to calibrate "error" as a function of DA metrics, we need to have an "error dataset", i.e., a large number of molecules predicted on the activity model, each with DA metrics and the absolute error of the prediction. Ideally one would like to build an activity model and generate predictions on a large number of molecules tested after the model was built. However, this is not necessarily practical. Our scheme[2,24] to approximate the ideal situation is to generate a large number of predictions by cross-validation.

DOI: 10.1021/acs.jcim.5b00110
J. Chem. Inf. Model. 2015, 55, 1098−1107

We assume that the activity model is built from data set T (in this application there are several versions, *T0.0*, *T0.3*, etc.) using method Q, descriptors D, and adjustable parameters P. The following procedure is used:

(1) Perform cross-validation:
  (a) Randomly assign a number n_t of molecules from T to be in the "training set". Molecules in T' that are not in the training set are in the "test set". Generate a QSAR activity model from the training set with method Q, descriptor(s) D, and parameters P.
  (b) Predict the molecules in the test set with the activity model. Make a note of the unsigned error (UE) for the prediction of each molecule i: $\mathrm{UE}(i) = |\mathrm{observed}(i) - \mathrm{predicted}(i)|$.
  (c) Calculate the DA metrics (see the next section) for each molecule i in the test set.
  (d) Repeat a–c, say, 10 times to build up the number of predictions.
(2) Pool the data for all of the predictions from all of the runs to form the error data set.

In step 1a, our practice is to vary n_t among the runs, with n_t ranging somewhere between $0.05N$ and $0.5N$, where N is the number of molecules in T or 10 000, whichever is smaller. Using a small n_t is less expensive computationally and allows for sampling of more molecules over a wider range for each DA metric. We showed in our previous work[24] that the size of the training set does not affect the relationship between the UE and the DA metrics, so it is valid to use n_t of various sizes where $n_t \ll N$. For very large data sets, it is more computationally tractable in step 1b to predict a large random sample (e.g., 5000 molecules) extracted from the test set rather than the entire test set. Typically we generate >50 000 pooled predictions in step 2. The same molecule may by chance occur in the test set of more than one cross-validation run, but these are treated as separate instances for the purposes of building an error model.

DA Metrics. In this work, we used the following DA metrics:

(1) TREE_SD is the standard deviation of the prediction for molecule M among the RF trees. If TREE_SD is small, i.e., the trees agree on the prediction, the overall prediction (the mean of the individual predictions) is expected to be more accurate. To calculate this metric, one needs to know the prediction of M on each RF tree.

(2) PREDICTED is the predicted activity value of M. Different ranges of activity may be better predicted than others. For many data sets, intermediate activity values have the highest errors, i.e., are the hardest to predict accurately. On the other hand, for data sets with very skewed activity distributions, the activity range with the least coverage may have the largest errors. Examples can be seen in Figure 2 of ref 18.

(3) SIMILARITYNEAREST1 is the similarity of M to the most similar molecule in the training set of the activity model.[2] It is expected that as SIMILARITYNEAREST1 goes up, the prediction of M is more likely to be accurate. To calculate this metric, one needs the AP descriptors of T and M.

(4) SIMILARITYNEAREST5 is the mean similarity of M to its five nearest neighbors. Since the metrics discussed next use the five nearest neighbors to M, it makes sense to include this metric.

(5) wRMSD1 is the weighted root-mean-square difference between the predicted activity of M and the observed activities of its five nearest neighbors in the training set:

$$\mathrm{wRMSD1} = \sqrt{\frac{\sum_k w(k)|\mathrm{observed}(k) - \mathrm{predicted}(M)|}{\sum_k w(k)}}$$

where k labels the neighbors and the weight $w(k)$ is taken as the similarity of M to k. This is based on the wRMSD defined by Keefer et al.[20] It is expected that as wRMSD1 decreases, i.e., the prediction of M agrees more with the observed activities of its neighbors, the prediction of M is likely to be more accurate. To calculate this metric, one needs the AP descriptors of T and M, the observed activities of the neighbors, and the prediction of M.

(6) wRMSD2 is the weighted root-mean-square difference between the predicted activities of the neighbors of M and the observed activities of the same neighbors:

$$\mathrm{wRMSD2} = \sqrt{\frac{\sum_k w(k)|\mathrm{observed}(k) - \mathrm{predicted}(k)|}{\sum_k w(k)}}$$

where $w(k)$ is the similarity of k to M. This is similar to the definition of Wood et al.[22] It is expected that as wRMSD2 decreases, i.e., the prediction of the neighbors of M gets more accurate, the prediction of M is likely to be more accurate. To calculate this metric, one needs the AP descriptors of T and M and the observed and predicted activities of the neighbors.

(7) NINTRAINING is the same as n_t. One might expect larger training sets to allow for more accurate predictions.

This is the same set of DA metrics as we used in our previous work.[24] It is not meant to be exhaustive but rather to be representative of the various types of metrics. TREE_SD is an example of a "bagged variance" metric. SIMILARITYNEAREST1 and -5 are examples of "distance to training set" metrics. wRMSD1 and wRMS2 are examples of "local error" metrics.

Construction of an Error Model from an Error Data Set (Section 4 in Figure 1). The major point of our previous paper[24] is that error models may be constructed from the error data set using random forest as the QSAR method. In this case, we use the R version of random forest (http://cran.r-project.org/web/packages/randomForest/randomForest.pdf) with some function of the UE as the "activity" and the seven DA metrics as the "descriptors". We construct error models with RF as regressions using all seven DA metrics or a subset of them. In our previous work, we discovered that the most consistent error models using all of the DA metrics, as decided by the leave-half-out cross-validated R^2 value (see Cross-Validation of an Error Model below), are obtained by using SQRT(UE) as the "activity" and running RF with a node size of 100, i.e., "leaves" in each tree with ≤100 molecules are not further split.

Generation of Prediction Uncertainty from the Error Model. In this study we are interested in the relative contribution of DA metrics to the prediction of SQRT(UE) in the error model. However, for completeness we will discuss how one goes from predictions on the error model for molecule M to prediction uncertainty for M, as described in our previous paper.[24] Although we calibrate and cross-validate the error model on individual SQRT(UE) values from the error data set, we do not claim to be able to predict the UE for M. Instead, we consider the prediction of M on the error model to represent

DOI: 10.1021/acs.jcim.5b00110
J. Chem. Inf. Model. 2015, 55, 1098–1107

the mean SQRT(UE) for molecules in the same region of DA metric space as *M*. For a Gaussian distribution of signed errors that is centered around zero, one can obtain the standard deviation of the Gaussian by scaling the mean UE by a factor of 1.26. If the "prediction uncertainty," i.e., the error bar on the predicted activity (which we called the "RMSE" in the previous paper), is considered equivalent to the "standard deviation of the Gaussian" and the "predicted UE" is considered to be the same as the "mean UE", then

$$\text{prediction uncertainty for } M = 1.26[\text{predicted SQRT(UE)}]^2$$

Descriptor Importance for an Error Model (Section 6 in Figure 1). In RF, "descriptor importance" is generated by seeing how much the accuracy of out-of-bag predictions are diminished when each individual descriptor is randomly assigned to a different molecule. This is a method of determining which descriptors are the most important in building a model. Here we are talking about the relative usefulness of the seven DA metrics to predict SQRT(UE) in the error model. We will refer to these as "DA metric importances" or "metric importances", to distinguish them from the importances of chemical descriptors on the activity model. Since the magnitude of any type of descriptor importance varies from data set to data set in the version of RF used here, one must normalize it when comparing data sets. To do this, we construct an error model with the SQRT(UE) values randomly reassigned to the molecules. The mean DA metric importance (over the seven metrics) for the error model built from scrambled data is called "BASELINE". This is the magnitude of DA metric importance that is equivalent to "noise". The normalized importance of a DA metric, say TREE_SD, for data set *j* will be adjusted relative to BASELINE:

$$\text{normalized metric importance}(j, \text{TREE\_SD}) = \text{metric importance}(\text{TREE\_SD}(j))/\text{BASELINE}(j)$$

All of the metric importances in this paper are calculated for error models built on all seven DA metrics.

Cross-Validation of an Error Model (Section 5 in Figure 1). Cross-validation is a standard QSAR approach for "validation". We have used cross-validation in several places. In section 1 of Figure 1, we can use cross-validation to see whether a self-consistent activity model can be built from the selected chemical descriptors. In sections 2 and 3 of Figure 1, we use cross-validation to generate an error data set. Here we can use cross-validation to see whether a self-consistent model for SQRT(UE) can be built from the DA metrics. In this case, we randomly select half of the SQRT(UE) values as a training set, and the other half becomes the test set. We make an error model from the training set and predict the SQRT(UE) values of the test set. One can use the mean R^2 over five cross-validation trials to measure the agreement of the predicted and observed SQRT(UE) values for the test set. We call this the cross-validated R^2 (CV-R^2) of the error model. This can be done with all seven DA metrics or a subset of them.

RESULTS

Self-Consistent Activity Models versus Self-Consistent Error Models. Data set sizes, MPSs, CV-R^2 values for both activity models and error models, and normalized descriptor importances for all subsets of the data sets in Table 1 are provided in the Supporting Information. Table 1 shows the CV-R^2 from the cross-validation of the activity model (Figure 1, sections 2 and 3) and the CV-R^2 of the error model (Figure 1, section 5) for *T0.0*. Consistent with our previous observation for diverse data sets,[24] whether one can build a self-consistent activity model for a given QSAR data set does not predict whether one can build a self-consistent error model. This holds true for the other subsets *T0.3* through *T0.7* (not shown).

Figure 4 shows the CV-R^2 of the error models as a function of the similarity cutoff and MPS using all of the DA metrics. It

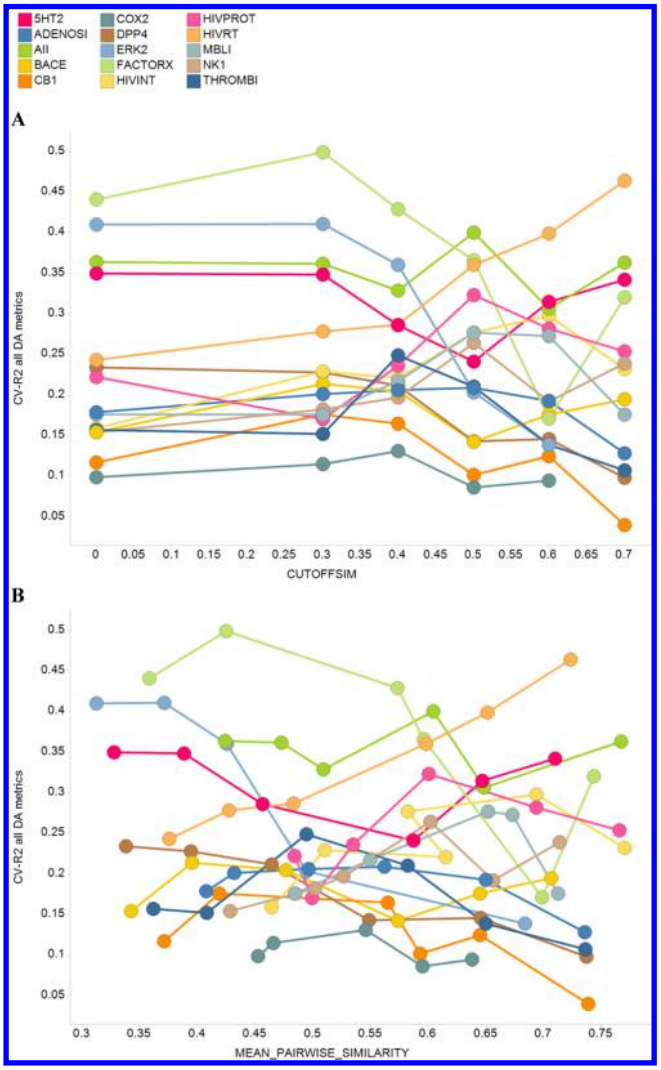


Figure 4. Cross-validated R^2 for the error model using all DA metrics vs (A) the cutoff similarity for the Butina algorithm or (B) mean pairwise similarity within each subset as the measure of diversity. Data sets are distinguished by color.

is clear that the relative order of data sets as measured by CV-R^2 may change as the diversity of the data sets changes. In this and subsequent figures, we have eliminated COX2 and ERK2 as data sets in *T0.7* because they have $N < 100$. This is too few molecules to get a credible sample of errors, and the CV-R^2 and normalized metric importances are not consistent with those for the rest of the data sets.

Review of Previous Results on Diverse Data Sets. To help provide context to the metric importance and CV-R^2 of the error model for the new data sets studied here, we show our previous results on 10 diverse data sets in Figures 5 and 6. Figure 5 is a modified version of Figure 6 (bottom) in ref 24.

DOI: 10.1021/acs.jcim.5b00110
J. Chem. Inf. Model. 2015, 55, 1098−1107

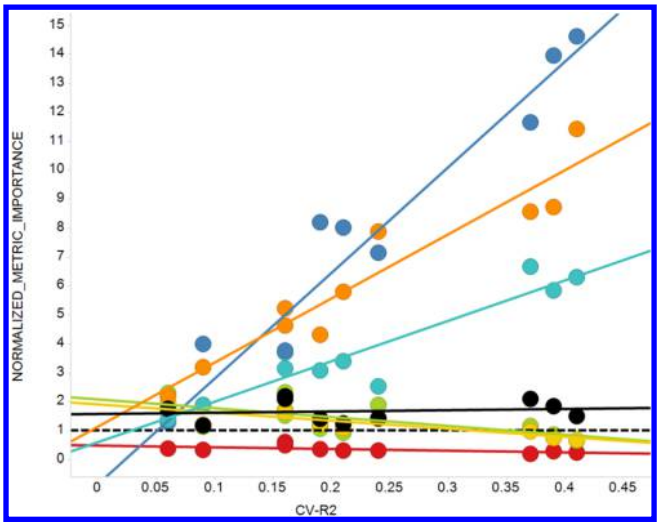


Figure 5. Normalized DA metric importances for error models against the cross-validated R^2 using all DA metrics. Metrics are represented by color: TREE_SD (orange), PREDICTED (blue), SIMILARITY-NEAREST1 (green), SIMILARITYNEAREST5 (yellow), wRMSD1 (cyan), wRMSD2 (black), and NINTRAINING (red). The horizontal dashed line at 1.0 represents the mean metric importance when SQRT(UE) is randomly assigned to the wrong molecule (i.e., BASELINE). Adapted from ref 24. Copyright 2013 American Chemical Society.

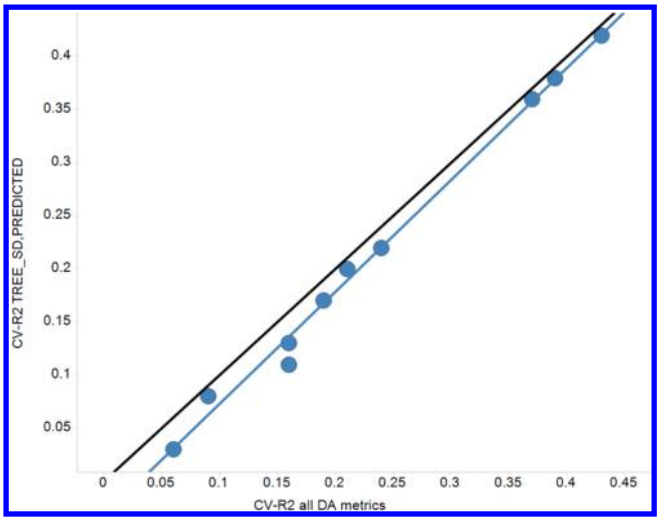


Figure 6. Cross-validated R^2 of the error model using TREE_SD, PREDICTED vs that using all of the DA metrics. Adapted from ref 24. Copyright 2013 American Chemical Society.

This plot displays the normalized metric importance as a function of the CV-R^2 of the error model using all seven DA metrics. Clearly, for many of the data sets, the CV-R^2 is low, and therefore, one cannot build error models for them given the DA metrics we used. For the data sets with reasonably high CV-R^2, i.e., where the largest normalized metric importance is several-fold higher than BASELINE, it is clear that PREDICTED (blue) and TREE_SD (orange) tend to be the most important DA metrics, with wRMSD1 (cyan) being third. The remaining metrics are never far above BASELINE, which means that they do not contribute significantly to the error model. Interestingly, the importance of TREE_SD, PREDICTED, and wRMSD1 is correlated with CV-R^2 for all of the data sets, but this is not true of the other metrics.

As only a few of the DA metrics appear to be important, one should be able to build an error model using only those DA metrics. Figure 6 is similar to Figure 7 (top) in ref 24. This shows a plot of the CV-R^2 of the error models using only the DA metrics TREE_SD and PREDICTED versus the CV-R^2 using all seven DA metrics. All of the points seem to be close to the diagonal, and they become even closer for high CV-R^2. We previously found that TREE_SD, PREDICTED, wRMSD1 is no closer to the diagonal than TREE_SD, PREDICTED and that TREE_SD and PREDICTED as single metrics are not as close to the diagonal. This implies that one needs only TREE_SD and PREDICTED in combination to make almost as good an error model for these diverse data sets as using all seven DA metrics.

DA Metric Importances as a Function of Diversity. Figure 7 shows the same type of plot as Figure 5, this time for the on-target data sets used in this study. Figure 7A is for *T0.0*, and Figure 7B is for *T0.7*. One observation is that the plot in Figure 7A (*T0.0*) is much like that in Figure 5. Again, TREE_SD, PREDICTED, and wRMSD1 seem to be correlated with CV-R^2, whereas the other DA metrics are not. This is not surprising since the *T0.0* data sets and the data sets in our

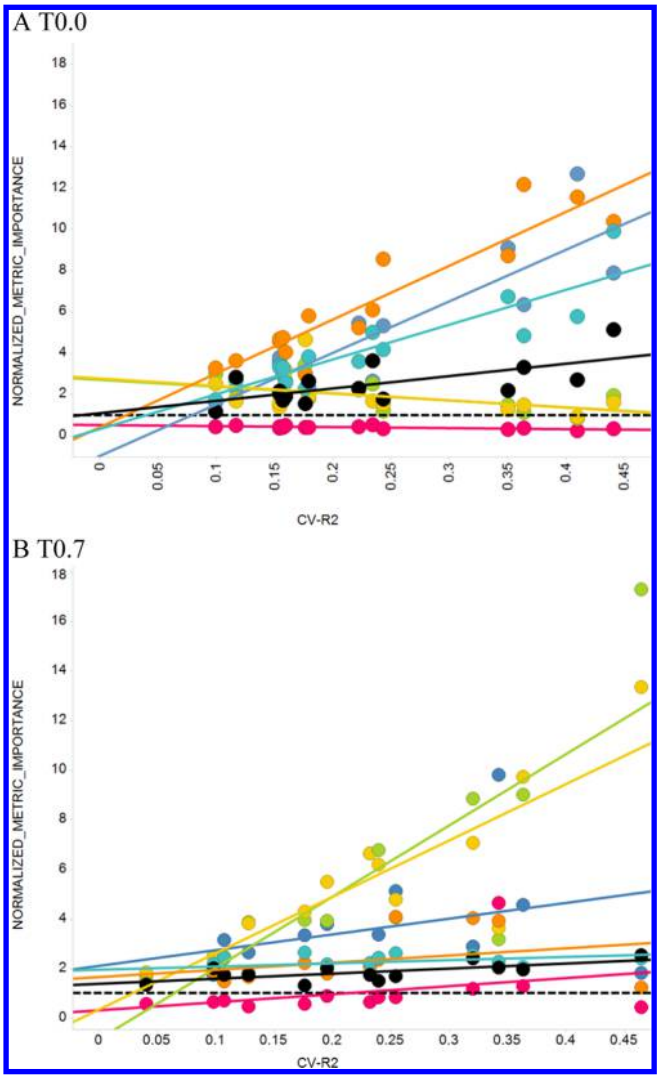


Figure 7. Normalized metric importance for *T0.0* and *T0.7* vs cross-validated R^2 of the error model. The DA metrics are distinguished by color using the same color scheme as in Figure 5.

DOI: 10.1021/acs.jcim.5b00110
J. Chem. Inf. Model. 2015, 55, 1098−1107

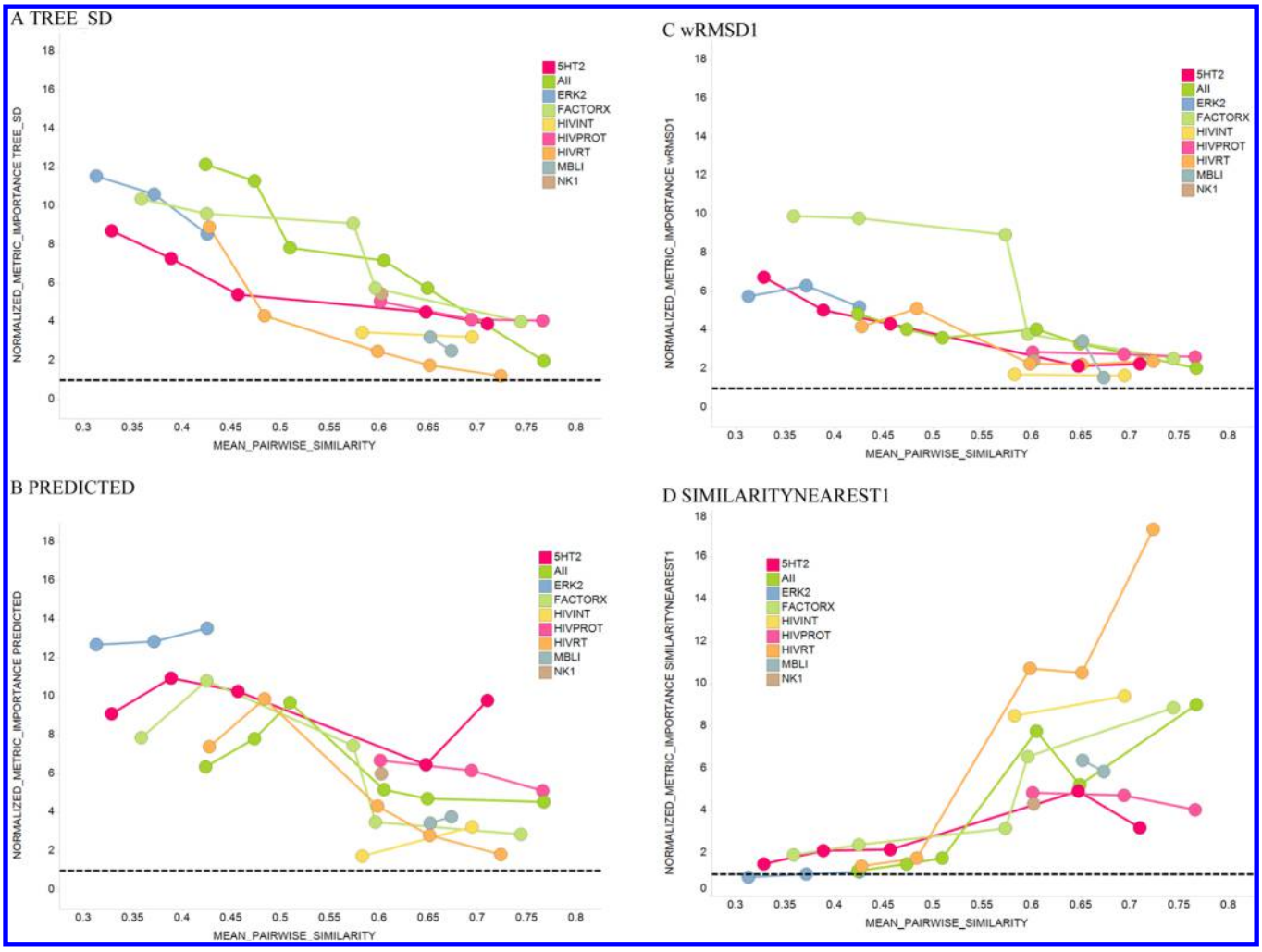


Figure 8. Normalized metric importance vs mean pairwise similarity for selected DA metrics. Data sets are distinguished by color. We show only the data set/T combinations for which the cross-validated R^2 is ≥0.25 using all of the DA metrics.

previous paper are all diverse. Figure 7B (*T0.7*) shows the extreme of less diversity. In this case, the most important DA metrics are SIMILARITYNEAREST1 and SIMILARITYNEAREST5, while TREE_SD and PREDICTED are much less important.

Figure 7 clearly shows that there is a transition with diversity, and this can be monitored by looking at graphs of normalized metric importance versus MPS. This is shown in Figure 8. We show only those data set/T combinations where a good error model can be built, i.e., those for which CV-$R^2 \geq 0.25$ using all seven DA metrics. (Since MPS tracks with the clustering cutoff, as shown in Figure 2, graphs of normalized DA importance vs the clustering cutoff tell the same story.) The importance of TREE_SD (Figure 8A) generally falls steadily with higher MPS (decreasing diversity), while SIMILARITYNEAREST1 (Figure 8D) rises with higher MPS. The results for SIMILARITYNEAREST5 (not shown) are almost identical to those for SIMILARITYNEAREST1. PREDICTED (Figure 8B) shows a more complicated trend where its importance starts on the high side, gets higher as MPS increases, and then falls as MPS is at a maximum. wRMSD1 (Figure 8C) also shows a decrease with increasing MPS; the trend is small except for FACTORXIA, where wRMSD1 starts out high. As expected, the normalized metric importances for the less significant error models (CV-R^2 < 0.25, not shown in Figure 8) are low over the entire range of MPS.

CV-R^2 for Subsets of Metrics as a Function of Diversity. Figure 9 shows the same type of plot as in Figure 6: CV-R^2 using only a subset of the DA metrics versus CV-R^2 using all seven DA metrics. Different clustering cutoffs are shown by color starting from *T0.0* (deep green) to *T0.7* (deep red). For TREE_SD, PREDICTED (Figure 9A), the points for cutoff = 0 (maximum diversity) hug the diagonal almost as closely as in Figure 5, but as the diversity decreases, the points fall away from the diagonal starting at *T0.5*, i.e., TREE_SD and PREDICTED account for less of the variation in SQRT(UE). Conversely, SIMILARITYNEAREST1 (Figure 9B) is far from the diagonal at *T0.0* and gets closer to the diagonal as the diversity decreases, meaning that SIMILARITYNEAREST1 accounts for much of the variation in SQRT(UE) at *T0.7*. However, these points do not get as close to the diagonal as TREE_SD, PREDICTED at low diversity. This means that a complete error model must take into account more DA metrics than just SIMILARITYNEAREST1. We would like calculate as few DA metrics as possible and still get as good a value of CV-R^2 as with all seven metrics. Since high-diversity data sets prefer TREE_SD, PREDICTED and low-diversity data sets prefer SIMILARITYNEAREST1, it is logical to try TREE_SD, PREDICTED, SIMILARITYNEAREST1. The points for this subset

DOI: 10.1021/acs.jcim.5b00110
J. Chem. Inf. Model. 2015, 55, 1098−1107

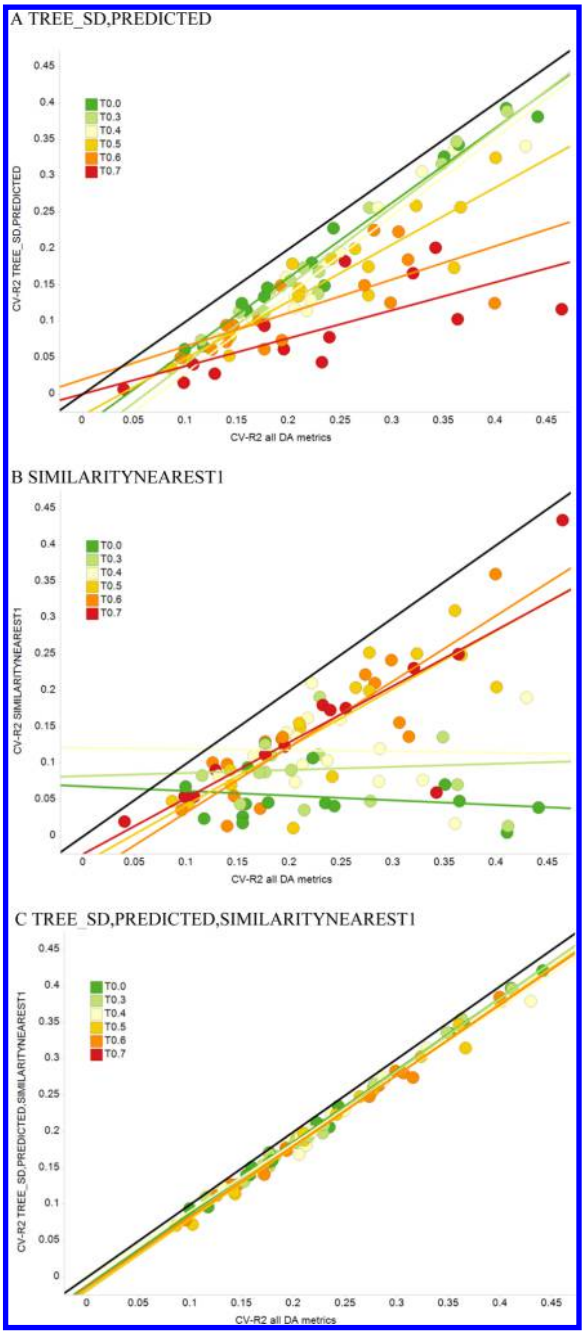


Figure 9. Cross-validated R^2 for the error model using a subset of DA metrics vs that using all of the DA metrics. Subsets of different diversity are shown by a color gradient going from deep green (*T0.0*) to deep red (*T0.7*).

of metrics hug the diagonal at any clustering cutoff (Figure 9C), so those three DA metrics seem sufficient.

DISCUSSION

Our previous paper looked at the relative importance of DA metrics in forming error models for diverse data sets for the purpose of generating prediction uncertainties. This paper examines the same question for less diverse data sets, and we find that nondiverse data sets have to be treated differently. Since we tried only a small sampling of DA metrics in both studies, we cannot rule out the possibility that some better DA metric, i.e., one that gets a better CV-R^2 in generating an error model, might exist. In that context, we note that in our laboratory we usually construct activity QSAR models with substructure descriptors, and we use similarity between molecules to define our DA metrics and to define diversity. Other laboratories routinely use property descriptors for their activity QSAR models, and for them it is more natural to use distances between molecules in the DA metrics and definition of diversity. We believe that our present results are robust for substructure descriptors, but we cannot say whether the relative importances of DA metrics will hold when the metrics are defined by property descriptors. However, the DA metrics TREE_SD and PREDICTED will probably be still be important in diverse data sets since they are not dependent on similarity or distances between molecules.

Of the DA metrics we tried, we found that with less diverse data sets (MPS $\geq$ 0.50), TREE_SD and PREDICTED become less important and metrics like SIMILARITYNEAREST1 become important. This makes perfect sense. We would expect that for a narrow data set (e.g., a single chemical series), the activity model should make a reasonably accurate prediction for *M* if *M* is an analogue in the series used to build the activity model. If *M* is not an analogue, there is no reason to expect the activity model to work, i.e., the predictions would be meaningless, and therefore, the mean UE would be large. Therefore, DA metrics that measure the similarity of *M* to the training set, such as SIMILARITYNEAREST1, should become important. In our previous paper,[24] we did speculate on the basis of similar reasoning that similarity to the training set might be important for less diverse data sets. Most chemists would intuit that similarity of *M* to the training set of a QSAR model would be good metric for prediction error for any training set, as we ourselves did in our first paper on domain applicability,[2] but this intuition is correct only for nondiverse data sets.

We can see from Figure 9C that the combination TREE_SD, PREDICTED, SIMILARITYNEAREST1 is almost as good at any level of diversity as using all seven DA metrics, and there is no reason to include the other metrics in making error models. Interestingly, these are the three metrics we declared "useful" for our 3DBINS method of generating prediction uncertainty.[18] There are some circumstances where we would not want to use all three metrics because of the issue of computational efficiency. While TREE_SD and PREDICTED are basically "free" with activity prediction from random forest and require no more additional calculations, SIMILARITYNEAREST1 requires calculation of the similarity of *M* to all compounds in the training set of the activity model. When training sets are large (>100 000 molecules), this can be a rate-limiting step. Generating the prediction uncertainty for *M* from an error model built for a large training set using TREE_SD, PREDICTED is about 10-fold faster than from an error model built using TREE_SD, PREDICTED, SIMILARITYNEAREST1. Fortunately, the TREE_SD, PREDICTED combination is adequate for predicting errors in diverse data sets, and in our environment the very largest data sets are for off-target activities and tend to be maximally diverse, i.e., they contain compounds from many chemical classes from many therapeutic areas.

We can form a general and practical rule: If a data set is diverse (MPS < 0.5) then TREE_SD, PREDICTED should be used to form the error model. Otherwise TREE_SD, PREDICTED, SIMILARITYNEAREST1 should be used. Less diverse data sets tend to be smaller, so calculating

DOI: 10.1021/acs.jcim.5b00110
J. Chem. Inf. Model. 2015, 55, 1098−1107

SIMILARITYNEAREST1 will not be computationally limiting. One can easily translate this diversity rule to another similarity definition. For example, a similarity of 0.5 in AP/Dice is equivalent to a similarity of 0.27 in ECFP4/Tanimoto.

■ ASSOCIATED CONTENT

Supporting Information

A text file containing a table of training set size, mean pairwise similarity, normalized descriptor importance, and CV-R^2 for multiple subsets. The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.5b00110.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: sheridan@merck.com.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was inspired by a conversation with Charlie (Zhenyu) Chang. The author thanks Joseph Shpungin for parallelizing random forest so that it can handle very large datasets. Dr. Andy Liaw suggested the SQRT transformation of unsigned error. The QSAR infrastructure used in this work depends on the MIX modeling infrastructure, and the author is grateful to other members of the MIX team. A large number of Merck biologists, over many years, generated the data for examples used in this paper.

■ REFERENCES

(1) Beck, B.; Breindl, A.; Clark, T. QM/NN QSPR models with error estimation: vapor pressure and logP. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1046−1051.

(2) Sheridan, R. P.; Feuston, B. P.; Maiorov, V. N.; Kearsley, S. K. Similarity to molecules in the training set is a good discriminator for prediction accuracy in QSAR. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1912−1928.

(3) Dimitrov, S.; Dimitrova, G.; Pavlov, T.; Dimitrova, N.; Patlewicz, G.; Niemela, J.; Mekenyan, O. A stepwise approach for defining the applicability domain of SAR and QSAR models. *J. Chem. Inf. Model.* **2005**, *45*, 839−849.

(4) Guha, R.; Jurs, P. C. Determining the validity of a QSAR model—a classification approach. *J. Chem. Inf. Model.* **2005**, *45*, 65−73.

(5) Tetko, I. V.; Bruneau, P.; Mewes, H.-W.; Rohrer, D. C.; Poda, G. I. Can we estimate the accuracy of ADME-Tox predictions? *Drug Discovery Today* **2006**, *11*, 700−707.

(6) Schroeter, T. S.; Schwaighofer, A.; Mika, S.; Laak, A. T.; Suelzle, D.; Ganzer, U.; Heinrich, N.; Muller, K.-R. Estimating the domain of applicability for machine learning QSAR models: a study on aqueous solubility of drug discovery molecules. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 485−498.

(7) Guha, R.; Van Drie, J. H. Structure−activity landscape index: quantifying activity cliffs. *J. Chem. Inf. Model.* **2008**, *48*, 646−658.

(8) Sprous, D. G. Fingerprint-based clustering applied to define a QSAR model use radius. *J. Mol. Graphics Modell.* **2008**, *27*, 225−232.

(9) Tetko, I. V.; Sushko, I.; Pandey, A. K.; Zhu, H.; Tropsha, A.; Papa, E.; Oberg, T.; Todeschini, R.; Fourches, D.; Varnek, A. Critical assessment of QSAR models of environmental toxicity against *Tetrahymena pyriformis*: focusing on applicability domain and overfitting by variable selection. *J. Chem. Inf. Model.* **2008**, *48*, 1733−1746.

(10) Weaver, S.; Gleeson, M. P. The importance of the domain of applicability in QSAR modeling. *J. Mol. Graphics Modell.* **2008**, *26*, 1315−1326.

(11) Dragos, H.; Marcou, G.; Varnek, A. Predicting the predictability: a unified approach to the applicability domain problem of QSAR models. *J. Chem. Inf. Model.* **2009**, *49*, 1762−1776.

(12) Kuhne, R.; Ebert, R.-E.; Schuurman, G. Chemical domain of QSAR models from atom-centered fragments. *J. Chem. Inf. Model.* **2009**, *49*, 2660−2669.

(13) Clark, R. D. DPRESS: localizing estimates of predictive uncertainty. *J. Cheminf.* **2009**, *1*, 11.

(14) Baskin, I. I.; Kireeva, N.; Varnek, A. The one-class classification approach to data description and to models applicability domain. *Mol. Inf.* **2010**, *29*, 581−587.

(15) Sushko, I.; Novotarskyi, S.; Körner, R.; Pandey, A. K.; Cherkasov, A.; Li, J.; Gramatica, P.; Hansen, K.; Schroeter, T.; Müller, K.-R.; Xi, L.; Liu, H.; Yao, X.; Öberg, T.; Hormozdiari, F.; Dao, P.; Sahinalp, C.; Todeschini, R.; Polishchuk, P.; Artemenko, A.; Kuz'min, V.; Martin, T. M.; Young, D. M.; Fourches, D.; Muratov, E.; Tropsha, A.; Baskin, I.; Horvath, D.; Marcou, G.; Muller, C.; Varnek, A.; Prokopenko, V. V.; Tetko, I. V. Applicability domain for classification problems: benchmarking of distance to models for Ames mutagenicity set. *J. Chem. Inf. Model.* **2010**, *50*, 2094−2111.

(16) Ellison, C. M.; Sherhod, R.; Cronin, M. T. D.; Enoch, S. J.; Madden, J. C.; Judson, P. N. Assessment of methods to define the applicability domain of structural alert models. *J. Chem. Inf. Model.* **2011**, *51*, 975−985.

(17) Soto, A. J.; Vazquez, G. E.; Strickert, M.; Ponzoni, I. Target-driven subspace mapping methods and their applicability domain estimation. *Mol. Inf.* **2011**, *30*, 779−789.

(18) Sheridan, R. P. Three useful dimensions for domain applicability in QSAR models using random forest. *J. Chem. Inf. Model.* **2012**, *52*, 814−823.

(19) Briesemeister, S.; Rahnenfuhrer, J.; Kohlbacker, O. No longer confidential: estimating the confidence of individual regression predictions. *PLoS One* **2012**, *7*, No. e48723.

(20) Keefer, C. E.; Kauffman, G. W.; Gupta, R. R. An interpretable, probability-based confidence metric for continuous QSAR models. *J. Chem. Inf. Model.* **2013**, *53*, 368−383.

(21) Gombar, V. K.; Hall, S. D. Quantitative structure−activity relationship models of clinical pharmacokinetics: clearance and volume of distribution. *J. Chem. Inf. Model.* **2013**, *53*, 948−957.

(22) Wood, D. J.; Carlsson, L.; Eklund, M.; Norinder, U.; Stålring, J. QSAR with experimental and predictive distributions: an information theoretic approach for assessing model quality. *J. Comput.-Aided Mol. Des.* **2013**, *27*, 203−219.

(23) Tetko, I. V.; Novotarskyi, S.; Sushko, I.; Ivanov, V.; Petrenko, A. E.; Dieden, R.; Lebon, F.; Mathieu, B. Development of dimethyl sulfoxide solubility models using 163000 molecules: using a domain applicability metric to select more reliable predictions. *J. Chem. Inf. Model.* **2013**, *53*, 1990−2000.

(24) Sheridan, R. P. Using random forest to model the domain applicability of another random forest model. *J. Chem. Inf. Model.* **2013**, *53*, 2837−2850.

(25) Gaspar, H. A.; Macrou, G.; Horvath, D.; Arault, A.; Lozano, S.; Vayer, P.; Varnek, A. Generative topological mapping-based classification models and their applicability domain: application to the biopharmaceutics drug disposition classification system (BDDCS). *J. Chem. Inf. Model.* **2013**, *53*, 3318−3325.

(26) Sahigara, F.; Ballabio, D.; Todeschini, R.; Consonni, V. Defining a *k*-nearest neighbors approach to assess the domain applicability of a QSAR model for reliable predictions. *J. Cheminf.* **2013**, *5*, No. 27.

(27) Toplak, M.; Močnik, R.; Polajnar, M.; Bosnić, Z.; Carlsson, L.; Hasselgren, C.; Demšar, J.; Boyer, S.; Zupan, B.; Stålring, J. Assessment of machine learning reliability methods for quantifying the applicability domain of QSAR regression models. *J. Chem. Inf. Model.* **2014**, *54*, 431−441.

(28) Liu, R.; Wallqvist, A. Merging applicability domains for in silico assessment of chemical mutagenicity. *J. Chem. Inf. Model.* **2014**, *54*, 793−800.

(29) Carrio, P.; Pinto, M.; Ecker, G.; Sanz, F.; Pastor, M. Applicability domain analysis (ADAN): a robust method for assessing

DOI: 10.1021/acs.jcim.5b00110
J. Chem. Inf. Model. 2015, 55, 1098−1107

the reliability of drug property predictions. *J. Chem. Inf. Model.* **2014**, *54*, 1500−1511.
(30) Noringer, 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.
(31) Kaneko, H.; Funatsu, K. Applicability domain based on ensemble learning in classification and regression analysis. *J. Chem. Inf. Model.* **2014**, *54*, 2469−2482.
(32) Sahlin, U.; Jeliazkova, N.; Oberg, T. Applicability domain dependent predictive uncertainty in QSAR regressions. *Mol. Inf.* **2014**, *33*, 26−35.
(33) Yan, J.; Zhu, W.-W.; Kong, B.; Lu, H.-B.; Yun, Y.-H.; Huang, J.-H.; Liang, Y.-Z. A combinatorial strategy of model disturbance and outlier comparison to define applicability domain in QSAR. *Mol. Inf.* **2014**, *33*, 503−513.
(34) Clark, R. D.; Lian, W.; Lee, A. C.; Lawless, M. S.; Frackiewicz, R.; Waldman, M. Using beta binomials to estimate classification uncertainty for ensemble models. *J. Cheminf.* **2014**, *6*, No. 34.
(35) Burden, F. R. Quantitative structure−activity relationship studies using Gaussian processes. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 830−835.
(36) Obrezanova, O.; Csanyi, G.; Gola, J. M. R.; Segall, M. D. Gaussian processes: a method for automatic QSAR modeling of ADME properties. *J. Chem. Inf. Model.* **2007**, *47*, 1847−1857.
(37) Oberezanova, O.; Segall, M. D. Gaussian processes for classification: QSAR modeling of ADMET and target activity. *J. Chem. Inf. Model.* **2010**, *50*, 1053−1061.
(38) Carhart, R. E.; Smith, D. H.; Ventkataraghavan, R. Atom pairs as molecular features in structure−activity studies: definition and application. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64−73.
(39) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.
(40) Butina, D. Unsupervised data base clustering based on Daylight's fingerprint and Tanimoto similarity: a fast and automated way to cluster small and large data sets. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 747−750.
(41) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. Random forest: a classification and regression tool for compound classification and QSAR modeling. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1947−1958.
(42) Kearsley, S. K.; Sallamack, S.; Fluder, E. M.; Andose, J. D.; Mosley, R. T.; Sheridan, R. P. Chemical similarity using physiochemical property descriptors. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 118−127.

DOI: 10.1021/acs.jcim.5b00110
J. Chem. Inf. Model. 2015, 55, 1098−1107