JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

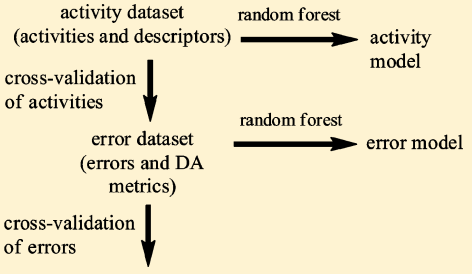
Using Random Forest To Model the Domain Applicability of Another Random Forest Model

Robert P. Sheridan*

Cheminformatics Department, Merck Research Laboratories, RY800-D133, 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. We will call this traditional type of QSAR model an "activity model". The activity model can be used to predict the activities of molecules not in the training set. A relatively new subfield for QSAR is domain applicability. The aim is to estimate the reliability of prediction of a specific molecule on a specific activity model. A number of different metrics have been proposed in the literature for this purpose. It is desirable to build a quantitative model of reliability against one or more of these metrics. We can call this an "error model". A previous publication from our laboratory (Sheridan *J. Chem. Inf. Model.*, **2012**, *52*, 814−823.) suggested the simultaneous use of three metrics would be more discriminating than any one metric. An error model could be built in the form of a three-dimensional set of bins. When the number of metrics exceeds three, however, the bin paradigm is not practical. An obvious solution for constructing an error model using multiple metrics is to use a QSAR method, in our case random forest. In this paper we demonstrate the usefulness of this paradigm, specifically for determining whether a useful error model can be built and which metrics are most useful for a given problem. For the ten data sets and for the seven metrics we examine here, it appears that it is possible to construct a useful error model using only two metrics (TREE_SD and PREDICTED). These do not require calculating similarities/distances between the molecules being predicted and the molecules used to build the activity model, which can be rate-limiting.



INTRODUCTION

In QSAR, a statistical model is generated from a training set of molecules (represented by chemical descriptors) and their biological activities. For the purpose of this paper we will call this traditional type of QSAR model an "activity model". The activity model can be used to predict the activities of molecules not in the training set. A relatively new subfield for QSAR is that of domain applicability (DA).[1−22] The idea is that some molecules are more or less likely to be predicted accurately with a given activity model and the goal is to find how to best distinguish the reliable predictions from the unreliable. This requires building another quantitative model, here called an "error model", of the likely errors of prediction on a specific activity model. A number of metrics have been demonstrated in the literature to be correlated with reliability of prediction (DA metrics). These come in a number of types: "distance/similarity to model", "bagged variance", "local error", etc.

The simplest type of error model expresses the reliability of the prediction as a function of a single DA metric and most of the DA literature uses this paradigm. In an earlier paper[18] we showed that more discrimination was possible using three metrics at once, and we devised a system of three-dimensional bins to act as a lookup table of prediction error. We called this the 3DBINS model. While easy to visualize, an error model in the form of bins is limited in the number of metrics it can handle simultaneously because it is hard to generate enough predictions in each bin for reasonable statistics. Moreover, the three metrics have to be chosen beforehand. Later, as new DA metrics appeared in the literature, we felt the need to test more than three metrics simultaneously. Ideally, one would like a paradigm for error models that can handle a larger number of candidate DA metrics and choose the important ones automatically. QSAR methods themselves (e.g., random forest, SVM, PLS, etc.) are in retrospect an obvious choice for this type of job. Where normally one would use "activity" in QSAR, one would use "prediction error". Instead of "chemical descriptors", one would use "DA metrics". In this paper we show that useful ideas normally used in the building of activity models (in particular cross-validation and descriptor importance) give insight into how errors are best modeled. We also show that these QSAR-based error models make reasonably quantitative estimates of errors for prospective predictions of activities.

METHODS

Overview. The overall use of activity and error models is shown in Figure 1. A conventional QSAR model, i.e. an activity model, is built using a QSAR method Q from an activity data set T, which is the combination of biological activities and chemical descriptors D (Section 1 in Figure 1). It is very common to use some flavor of cross-validation to decide which combination of descriptors and method, plus any adjustable parameters, give the

Received: August 15, 2013
Published: October 23, 2013

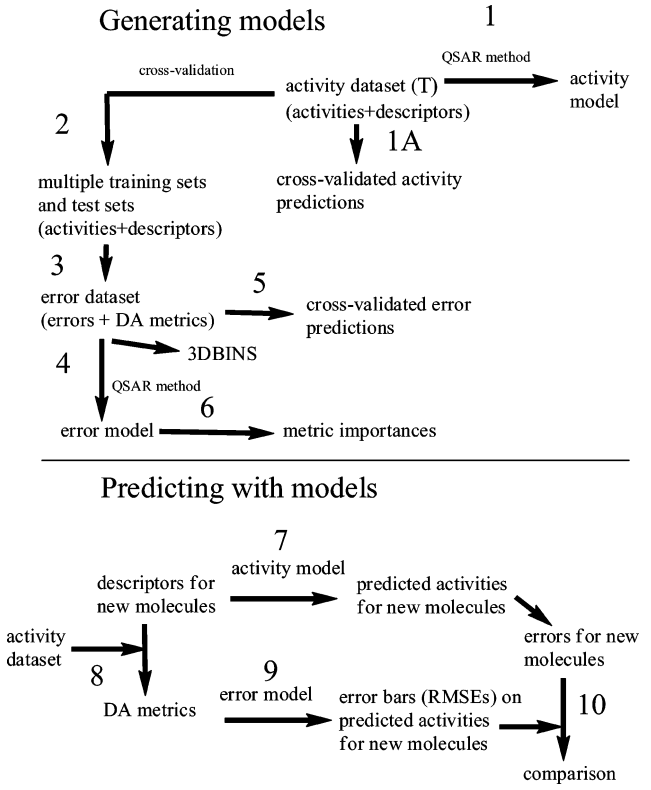


Figure 1. Scheme for use of activity models and error models.

best activity model (Section 1A). For the purposes of this paper we will assume that it has been decided to use descriptors *D*, method *Q*, and adjustable parameters *P* for the activity model.

To build an error model associated with the activity model one must start with an error data set, a set of predictions on the activity model with one or more DA metrics associated with each prediction. Our proposed method[2,18] of approximating this is 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 (Section 7). One generates the DA metrics for *M* using the activity model and/or the chemical descriptors for *T* (Section 8). To make an error prediction (i.e., the reliability) for *M*, one scores the DA metrics for *M* against the error model (Section 9).

Details follow:

QSAR Methods and Descriptors for the Activity Model (Section 1 in Figure 1). There is a large number of useful QSAR methods. Our method of choice, and one that we will use here, is random forest (RF).[23] RF is an ensemble recursive partitioning method where each recursive partitioning "tree" is generated from a bagged sample of molecules, with a random subset of descriptors used at each branching of each tree. Typically we generate 100 trees; adding further trees does not improve prediction accuracy. Useful features of RF are that it produces cross-validated predictions at least as good as most other QSAR methods, it can handle nonlinear relationships, one does not have to do descriptor selection to obtain good results, coupling between descriptors is easily handled, and predictions appear robust to changes in the adjustable parameters. RF can do regressions or classifications. RF is very attractive for DA purposes because we get a "bagged variance" DA metric with no extra work.

One nonstandard practice we apply for RF is "prediction rescaling". Because of how predictions are made in RF regression (by the average value at terminal nodes), sometimes the range of predictions in RF is compressed relative to the range of observations by 10−20%. In effect, the higher activity values are systematically predicted too low and the lower activity values are predicted too high. To remedy this, when the activity model is built, we determine the best linear relationship that will transform the self-fit predictions of molecules to their observations. We store this relationship with the model and apply it to the raw predictions from the model to get a rescaled prediction. When we talk about predictions of an RF model in this paper, we will mean the rescaled predictions, both for activity models and error models. While not changing the correlation between predictions and observations, the rescaling helps get the numerical value of the predictions more correct, particularly for the highest and lowest part of the activity range.

There is also a large number of possible chemical descriptors that can be used for activity models. In our hands the union of the Carhart atom pairs (AP)[24] and a donor−acceptor pair (DP), called "BP" in Kearsley et al.,[25] give us the most accurate cross-validated predictions. Both descriptors are of the form

$$\text{Atom type i} - (\text{distance in bonds}) - \text{Atom type j}$$

For AP, atom type includes the element, number of nonhydrogen neighbors, and number of pi electrons. For DP, atom type is one of seven (cation, anion, neutral donor, neutral acceptor, polar, hydrophobe, and other).

Definition of Errors. Figure 2 can help understand the definitions of errors. We have a set of molecules with an observed activity and a predicted activity. One common metric for the "goodness of prediction" in QSAR is the R^2, which captures the correlation between predicted and observed. The other common metric is the numerical difference between observed and

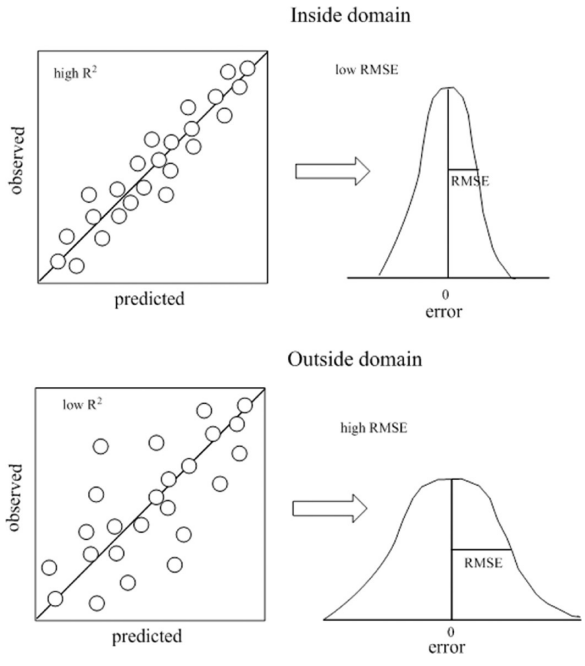


Figure 2. Cartoon of agreement between predicted and observed activities in two types of situations: for molecules in the domain of the training set and outside the domain. Agreement between predicted and observed values can be measured by R^2 or root-mean-square error.

predicted. We will define the "error" for molecule i as observed minus predicted. Usually one assumes that positive and negative errors are equally likely, and therefore one is concerned only with the unsigned error (UE) of molecule i:

$$\mathrm{UE(i)} = |\mathrm{observed(i)} - \mathrm{predicted(i)}|$$

In QSAR one usually uses the root-mean-square error (RMSE) over a set of molecules as a measure of the "goodness of prediction". One can imagine a histogram of errors for a set of molecules. This is usually close to a Gaussian distribution with the mean close to zero and the RMSE being equal to the standard deviation of the Gaussian.

Let us assume we can find a DA metric that can effectively separate reliable from unreliable predictions. For the purposes of this discussion let us say the DA metric is similarity to the nearest compound in the training set, and compounds with similarity >0.9 will be more reliably predicted ("inside the domain"). In a graph of observed vs predicted, the molecules will be found close to the diagonal (i.e., the UE is small), the R^2 will be high and the RMSE small. In contrast, molecules with a similarity <0.2 ("outside the domain"), the molecules will not be so close to the diagonal, the R^2 will be low and the RMSE high. Note that even "outside the domain" the most common error is near zero; what changes is the range of likely errors. In the field of DA, the idea is to assign "error bar" on a prediction for a specific molecule *M*, and this can be in units of RMSE; a smaller RMSE means a higher "reliability". Of course, RMSE is not meaningful for the prediction of a single molecule, and we cannot estimate the actual UE for *M*, which can be known only after the activity for *M* is actually measured. We are trying to estimate the "error bar" for molecules with a similarity to the training set like that of *M*. Above we used similarity to the data set as one DA metric, but the concept can be extended to a function including any number of metrics.

We should note that the magnitude of the "error bar" in terms of RMSE not only includes the uncertainty due to domain applicability but also includes the experimental uncertainty in the activity.

DA Metrics. For this work we will use the DA metrics below. This is not meant to be an exhaustive list by any means, just representative of the types of metrics in the literature.

1. TREE_SD. This is the standard deviation of the prediction for molecule *M* among the 100 RF trees. The expectation is that if TREE_SD is small, on the average UE will be smaller. That is, if the trees agree on the prediction, the prediction is likely to be more accurate. This is an example of a "variation among bagged subsets" DA metric. To calculate this metric one needs to know the prediction of *M* on each RF tree.

2. PREDICTED, i.e. the predicted activity value of *M*. Our earlier work[18] showed that some ranges of activity are more easily predicted than others, and this is a strong effect. For many data sets, the mean UE tends to be higher in the middle of the activity range, but in some data sets the highest UE is at the low or high activity range.

The next DA metrics, examples of "similarity/distance to model" metrics, depend on the concept of similarity and the concept of nearest neighbors in the training set, i.e. the compounds most similar to M. We will follow our previous practice of defining similarity based on the AP descriptor and the Dice similarity index. Using an alternative definition of similarity such as the ECFP4 descriptor and the Tanimoto index gives very similar results.

3. SIMILARITYNEAREST1. This 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 average UE will go down. That is, the prediction of *M* is more likely to be accurate if it is close to at least one molecule in the training set. To calculate this metric one needs AP descriptors of *T* and of *M*.

4. SIMILARITYNEAREST5. Since the metrics discussed next use five nearest neighbors to *M* it makes sense to include this, which is the mean similarity of *M* to those neighbors.

Then next metrics are of the "local error" type:

5. wRMSD1 is the weighted root-mean-square difference between the predicted activity of *M* and the observed activity of its five nearest neighbors

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

where *k* goes over the neighbors. We use $w(k)$ = the similarity of *M* to *k*. This is very close to the wRMSD defined by Keefer et al.,[20] using similarity instead of distance to define neighbors. It is expected that when wRMSD1 goes up, UE on the average will go up. That is, if the prediction of *M* agrees more with the observed values 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 activity of the neighbors of *M* and the observed activity of the same neighbors

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

where $w(k)$ is the similarity of *k* to *M*. This is close to the definition of Wood et al.[20] except using similarity instead of distance to define neighbors. Similar indices have been proposed by others, for example in refs 13 and 19. It is expected that when wRMSD2 goes up, the average UE will go up. That is, if the neighbor(s) of *M* are accurately predicted, *M* is more likely to be accurately predicted. To calculate this metric one needs the AP descriptors of *T* and *M* and the observed and predicted activities of the neighbors. It should be noted that in Wood et al. the predicted(k) is from a cross-validated prediction of the training set, whereas here we use the self-fit prediction of *k* from the building of the activity model from *T*. We have found that the errors from the self-fit prediction track well with the errors from cross-validated prediction, albeit the magnitude of the first is typically 2-fold smaller. Thus, the wRMSD2 using the self-fit prediction should be useful, and the extra work of "double-loop" cross-validation is avoided.

7. NINTRAINING. This is the number of the molecules in the training set.

As one might expect SIMILARITYNEAREST1 and SIMILARITYNEAREST5 are correlated with each other. NINTRAINING tends not to be significantly correlated with other metrics. Other pairwise correlations vary in magnitude, and even sign, from data set to data set.

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 a large number of molecules predicted on the activity model with the DA metrics of each molecule.

dx.doi.org/10.1021/ci400482e | J. Chem. Inf. Model. 2013, 53, 2837−2850

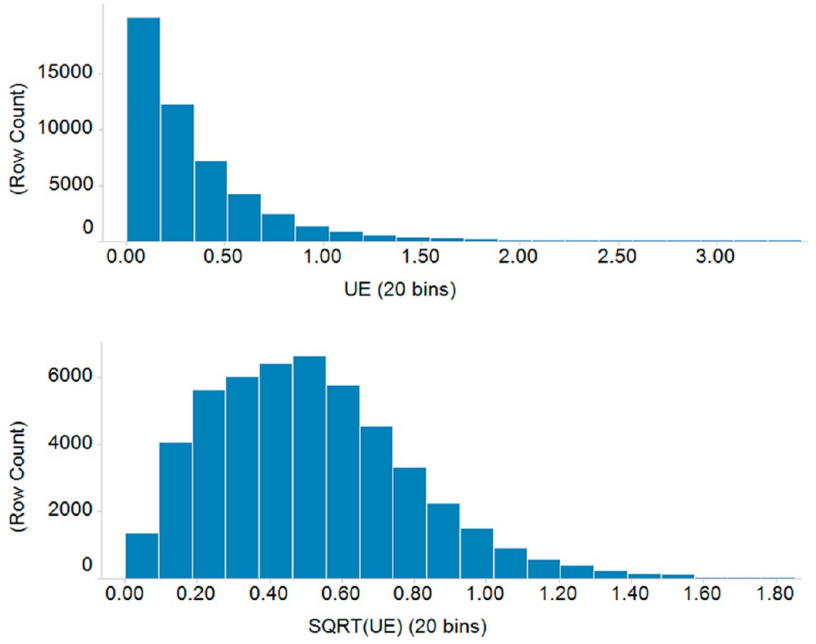


Figure 3. Distribution of unsigned errors from the cross-validation of the 3A4 example (Section 3 from Figure 1). Top: the raw unsigned error. Bottom: the square root of the unsigned error.

In an ideal world the predictions would be prospective. One would generate the activity model using all available data, wait for enough diverse molecules not already in the model to be tested for the same specific activity, predict those molecules with the activity model, calculate the DA metrics for the molecules, and process the information into an error model. This is difficult in practice, however. Our original suggestion[2] to approximate the ideal situation was to generate a large number of predictions by cross-validation.

Assume the activity model is built from data set T using method Q, descriptors D, and adjustable parameters P.

1. Perform cross-validation:

   a. Randomly assign a number n_t of molecules from T to be in the "training set". Whatever molecules remain in T are in the "test set". Generate a QSAR model from the training set with method Q, descriptors D, and parameters P.
   b. Predict the molecules in the test set with the model. Make a note of the UE for each prediction.
   c. Calculate the DA metrics for each molecule in the test set.
   d. Repeat a-c, say, 10 times to build up the number of predictions.

2. Pool the data for all predictions from all the runs to form the error data set.

Our current practice for step 1a 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. Having small n_t is less expensive computationally and allows for sampling more molecules over a wider range for each DA metric. For instance, when n_t is small there are more examples where SIMILARITYNEAREST1 is low. The relationship of error vs DA metrics should not be sensitive to n_t, otherwise the idea of using cross-validation with $n_t \ll N$ to generate an error data set would not be valid. The lack of sensitivity is easy to verify with a single DA metric. One can see, for instance, that the placement of the curve of RMSE vs SIMILARITYNEAREST1 does not change with n_t, as shown in Sheridan et al.[2] Later in the Results section we will show that this is true when we use several DA metrics simultaneously.

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 predict the entire test set.

Typically we generate >50,000 pooled predictions in step 2. The same molecule may by chance occur in more than one cross-validation run. For the purposes of constructing an error data set they are treated as separate instances.

Constructing an Error Model from an Error Data Set (Section 4 in Figure 1). The major point of this paper is that error models may be constructed from the error data set by a QSAR method. Some function of UE will be the "activity" and the DA metrics will be the "descriptors". One complication is that, if we assume the error bars on the prediction of M need to be in units of RMSE, we cannot fit RMSE directly with the QSAR method since RMSE is not defined for a single molecule. However, we can fit UE, which is defined for each M in the error data set. As will be shown below, we can rescale predicted UE from the error model to units of RMSE.

The same features that make RF attractive for activity models also make it attractive for error models. One specific feature, compared to methods such as PLS, is that it can handle nonlinear relationships; this is useful because, as mentioned above, UE is often a parabolic function of the DA metric PREDICTED. We construct error models with RF using all 7 DA metrics. All error models are built as regressions. Again we use 100 trees.

We investigated two variations in an attempt to optimize the fit of UE using RF. Besides fitting raw UE, we also tried SQRT(UE). The distribution of UE is almost always skewed toward lower values, and SQRT(UE) makes the distribution more normal and therefore possibly easier for any QSAR method to fit. An example is shown in Figure 3. Nodesize is the size of the node in a recursive partitioning tree that will no longer be split. Higher nodesizes imply a smoother fit. Our experience with RF suggests that if there are few parameters, better fits can be found with

dx.doi.org/10.1021/ci400482e | J. Chem. Inf. Model. 2013, 53, 2837−2850

Table 1. Data Sets

name	description	source	N	CV-R^2 for activity model	CV-R^2 for error model SQRT(UE) nodesize = 100
1A2	−log(IC50) for CYP 1A2 inhibition	Pubchem AID 1851 refs 27, 28	13,243	0.47	0.37
3A4	−log(IC50) for CYP 3A4 inhibition	in-house	140,575	0.46	0.39
AIDS	−log(EC50) for protection of cells against HIV	ref 29	38,870	0.15	0.19
FACTORX	−log(IC50) or −log(K_i) Inhibition of human factorX	Chembl version 8 TID=194 ref 30	4,784	0.65	0.16
HERG	−log(IC50) for binding to hERG channel	in-house	265,678	0.33	0.09
HPLC_LOGD	LOGD measured by retention time on HPLC	in-house	250,575	0.57	0.06
NK1	−log(IC50) for binding to substance P receptor	in-house	13,482	0.67	0.16
PGP	log(BA/AB) for active transport by p-glycoprotein	in-house	11,106	0.54	0.24
PXR	induction of pregnane X receptor relative to rifampicin	in-house	139,331	0.36	0.21
Pyruvate_kinase	−log(IC50) of inhibition of Bacillus pyruvate kinase	Pubchem AID 361 ref 31	51,441	0.06	0.41

larger nodesizes. We tried several different nodesizes: 5 (the default for regression problems), 25, 50, 100, 200, 400.

Cross-Validating an Error Model (Section 5 in Figure 1). Cross-validation is a standard QSAR approach for "validation". In the case of error models one may rationally choose between UE and SQRT(UE) and the nodesizes by seeing which combination is most internally consistent. In this case we randomly select half of the errors as a training set and the other half becomes the test set. We make an error model with the training set and predict the test set. One can use the mean R^2 over five cross-validation trials to measure the agreement of predicted UE and observed UE of the test set.

Descriptor Importance for an Error Model (Section 6 in Figure 1). Through descriptor importance we can estimate the relative usefulness of DA metrics to predict UE for a given error data set. In RF, descriptor importance is generated by seeing how much the accuracy of out-of-bag predictions is diminished when each individual descriptor is randomly assigned to a different molecule. Since the magnitude of the descriptor importance varies from data set to data set, one must normalize it when comparing data sets. To do this we construct an RF error model with the UE or SQRT(UE) randomly assigned to the wrong molecules. The mean importance over all DA metrics for that situation is defined as BASELINE, the magnitude of descriptor importance that signifies "noise". The normalized importance of a DA metric for data set j, say TREE_SD, will be

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

3DBINS Error Model (Section 4 in Figure 1). As an alternative to an error model built with random forest, we also create an error model using the 3DBINS paradigm described in our previous work.[18] 3DBINS uses the DA metrics TREE_SD, PREDICTED, and SIMILARITYNEAREST1. Each dimension is divided into 9 intervals; this forms 273 bins. For each bin (for example, $0.25 < \text{TREE\_SD} \leq 0.50$, $0.75 < \text{PREDICTED} \leq 1.0$, $0.60 < \text{SIMILARITYNEAREST1} \leq 0.7$) one can calculate the RMSE for the predictions that fall in that bin. The set of bins acts as a lookup table: For the prediction of a new molecule M, one finds the bin to which *M* is closest in the three DA metrics. One uses the RMSE of the closest bin as the "error bar" for the activity prediction of M.

Comparing the Observed and Predicted UE and RMSE in Prospective Prediction (Sections 7−10 in Figure 1). Consider a set of molecules assayed for the appropriate activity after the data set *T* was compiled. Predictions of these molecules will be "prospective". One generates an activity prediction of each new molecule *M* with the activity model (Section 7). Given the observed and predicted activity of *M*, one then has an observed UE or observed SQRT(UE). One then generates the DA metrics for this set of molecules (Section 8). Given the DA metrics, one uses the error model to produce "predicted UE" and/or "predicted SQRT(UE)" for each *M* (Section 9). One can compare the predicted vs observed UE or SQRT(UE) (Section 10).

Ultimately one must compare the predicted RMSE to the observed RMSE. For Gaussian distributions, RMSE = ~1.26(mean UE). Therefore the predicted RMSE of *M* from the error model will be 1.26(predicted UE) or 1.26(predicted SQRT(UE))^2 depending on whether the error model was made from UE or SQRT(UE), respectively.

Observed RMSE can be calculated only for a set of molecules, so assigning an observed RMSE to *M* involves averaging over molecules that closely resemble *M* in their predicted RMSE. Consider the set of molecules with predicted RMSE within $\pm d$ of the predicted RMSE of *M*, where *d* is one-fiftieth of the range in 1.26(observed UE). The observed RMSE for *M* is the root-mean-square of the observed UE for those molecules.

Data Sets. We will show as examples 10 QSAR data sets listed in Table 1. We tried to include target-specific data sets and ADME data sets of various sizes (~1000 to >250,000 molecules) from various sources. Some are from the open literature included for the purposes of "reproducibility", while others are proprietary data sets from Merck. It is necessary to include some proprietary data in this study. While it is not hard nowadays to find realistically large (>10,000 molecules) data sets from PubChem,[26] these are limited to high-throughput screens and/or confirmation data. Also, some of our later tests require prospective validation, where the activity of molecules is measured after the model is built, and dates of testing are easily available in-house but nearly impossible to find in publicly available data sets.

Some of these data sets have a substantial fraction of "qualified data" for example, "IC50 >30 μM". Most off-the-shelf QSAR methods, including the implementations of RF we use here, do

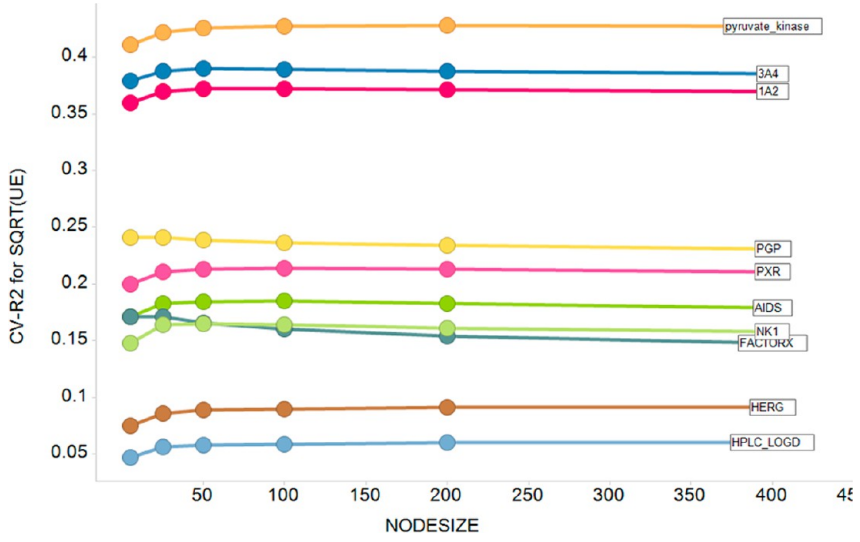


Figure 4. Cross-validated R^2 for error models as a function of nodesize for SQRT(UE). Each line represents a data set.

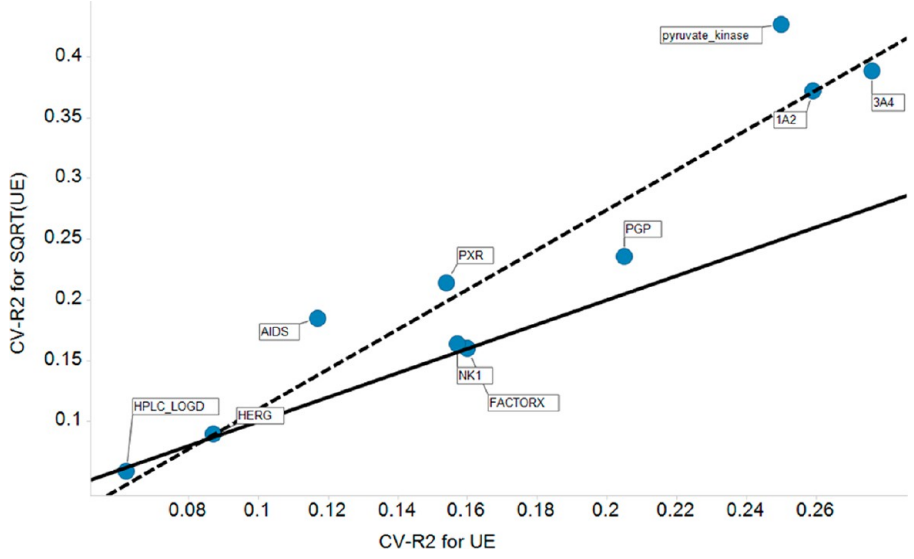


Figure 5. R^2 of cross-validated prediction for SQRT(UE) vs UE using nodesize = 100. The diagonal is represented as a solid line, and the best linear fit through the data is represented by a dashed line.

not treat qualified data as anything but a definite value, i.e. >30 μM would be treated as 30 μM (or $-\log(IC50) = -4.5$ in units of M). The activity models for these data sets are all regressions.

Some in-house data sets (3A4, HERG, HPLC_LOGD, PXR) are from high-throughput assays, and data has accumulated for tens of thousands of molecules since the construction of the data sets in Table 1. Predicting the activities and the errors for these molecules would be an example of "prospective prediction" since the compounds were tested at a later date than all the molecules in the activity model. We selected a random sample of 10,000 new molecules each for 3A4, HERG, HPLC_LOGD, and PXR. The PGP assay is not as high-throughput, and we have only ~3,000 molecules for prospective prediction.

RESULTS

Cross-Validation of Activity Data Sets. The R^2 for the agreement of predicted and observed activity from the cross-validation in Section 2 of Figure 1 is listed in Table 1. Clearly some activities are easier to predict than others, NK1 and FACTORX being among the best, and AIDS and pyruvate_kinase being the worst.

Cross-Validation of Error Data Sets. Cross-validation (Section 5 in Figure 1) was used to decide whether UE or SQRT(UE) is the proper transformation and to decide which nodesize to use for the error model. In Figure 4 we see that the cross-validated R^2 tends to rise with increasing nodesize but has effectively plateaued at nodesize = 100 for SQRT(UE). The same is true for UE (not shown). In Figure 5 we see that R^2 is clearly systematically higher for SQRT(UE) than UE, becoming more favored as R^2 increases. This is not surprising, and at least partly a mathematical consequence of the SQRT transformation moving the distribution of values of observed UE away from zero; this makes the relationship between observed UE and predicted UE less of a lower-diagonal relationship and more of a linear relationship, which is assumed when we calculate R^2.

From now on, all discussion of RF error models will assume SQRT(UE) and nodesize = 100.

Cross-validated R^2 for nodesize = 100 and SQRT(UE) is listed in Table 1. Clearly some error models are better than others, with

dx.doi.org/10.1021/ci400482e | J. Chem. Inf. Model. 2013, 53, 2837−2850

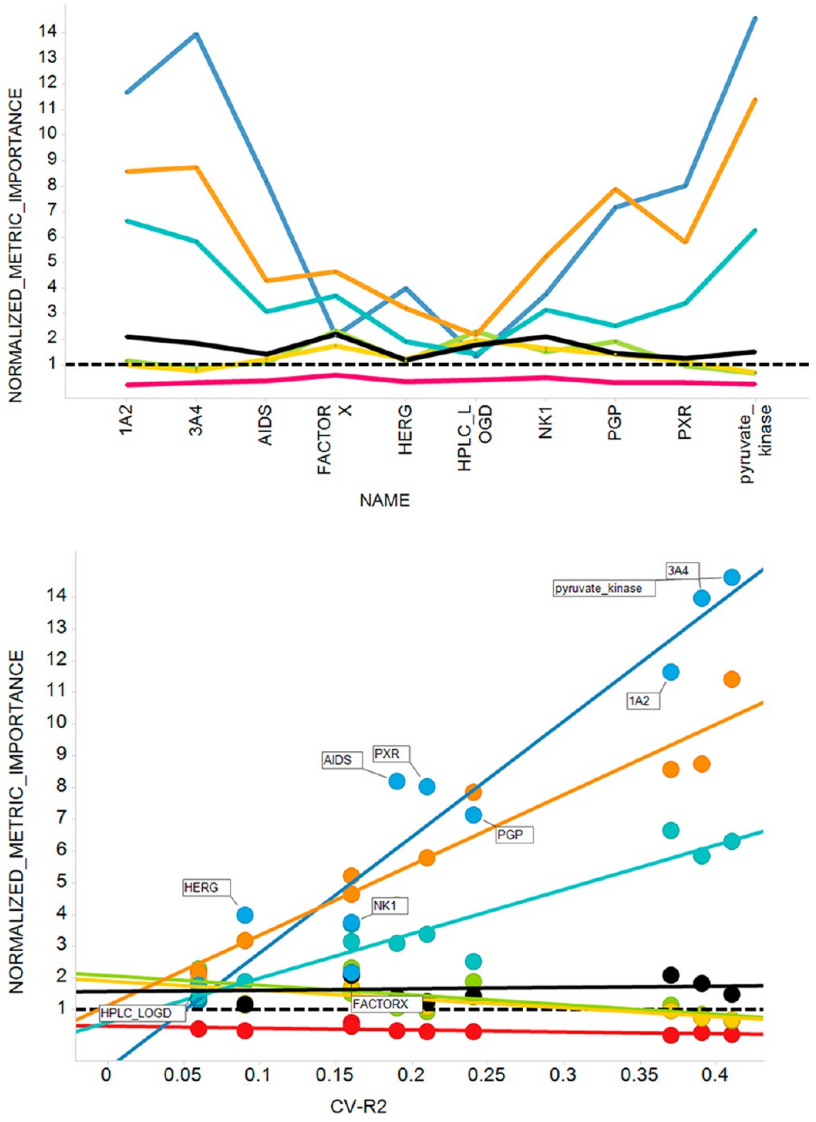


Figure 6. Normalized metric importances for error models fitting SQRT(UE) using nodesize = 100. Top: Arranged by the names of the data sets. Bottom: Plotted against the cross-validated R^2 for the error model. Metrics are represented by color: TREE_SD (orange), PREDICTED (blue), SIMILARITYNEAREST1 (green), SIMILARITYNEAREST5 (yellow), wRMSD1 (cyan), wRMSD2 (black), NINTRAINING (red). The horizontal dashed line represents the mean importance when the SQRT(UE) is randomly assigned to the wrong molecule, i.e. BASELINE.

pyruvate_kinase being the best and HERG and HPLC_LOGD being the worst. One might expect that data sets with good activity models might have poor error models since the magnitude of the error is small, and that the reverse would be also be true. However, there is no obvious relationship between the cross-validated R^2 of the activity model and the cross-validated R^2 of the error model. That is, one can have a good activity model with a good (e.g., 3A4), medium (e.g., PGP), or poor (e.g., HPLC_LOGD) error model. On the other hand, one may have a very poor activity model and a good (e.g., pyruvate_kinase) or poor (e.g., AIDS) error model. We do not have an explanation for this at present.

Importance of DA Metrics. Figure 6 (top) shows the normalized metric importance for the error models of the 10 data sets (Section 6 in Figure 1). Consistent with our previous results,[18] we find that PREDICTED (blue) and TREE_SD (orange) tend to be the most important DA metrics for most data sets. The next most important metric tends to be wRMSD1 (cyan). The remaining metrics are never far above BASELINE (the dashed line at normalized metric importance = 1). Gratifyingly, NTRAINING (red) is below BASELINE for every data set; the assumption that the size of the training set in the cross-validation of activity models (Section 2 of Figure 1) does not systematically affect the estimation of UE appears to be true. Figure 6 (bottom) shows the same data as a function of the cross-validated R^2 of the error model (as listed in Table 1). It is clear that the overall cross-validated R^2 is correlated with the importance of TREE_SD, PREDICTED, and wRMSD1 but not with any of the other metrics.

Cross-Validation of Error Models with Subsets of Metrics. Given that some DA metrics are less important we can in principle omit them in building error models. Figure 7 (top) shows the cross-validated R^2 of the error models (Section 5 in Figure 1) using three or fewer metrics plotted against the cross-validated R^2 of the error model using all seven metrics. In that figure, models with fewer metrics that are closest to the diagonal are best approximations for the full model. RF error models with TREE_SD, PREDICTED, wRMS1 (red), and

dx.doi.org/10.1021/ci400482e | J. Chem. Inf. Model. 2013, 53, 2837−2850

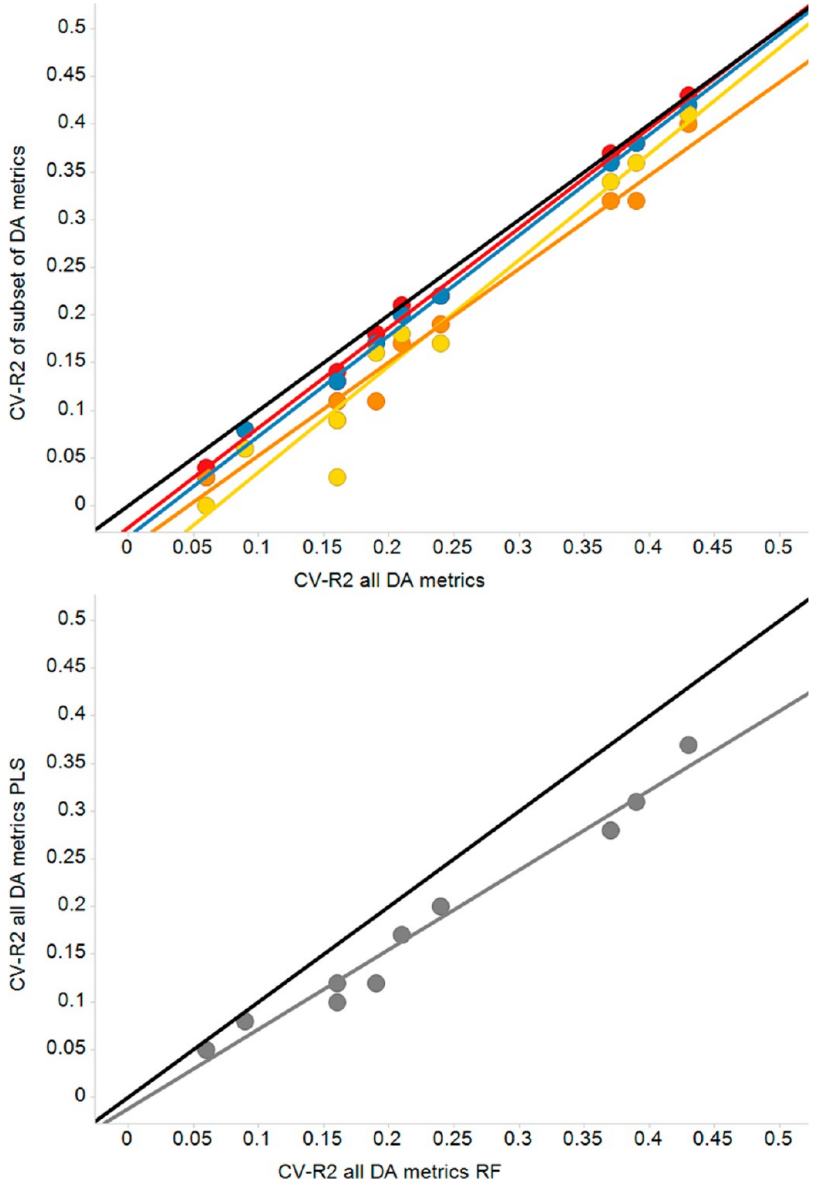


Figure 7. Cross-validated R^2 of the error model for SQRT(UE) and nodesize = 100 (for RF). Top: RF error models using subsets of DA metrics vs all metrics. A perfect approximation to using all metrics would lie against the diagonal (black line). Models are distinguished by color: TREE_SD,PREDICTED,wRMSD1 (red), TREE_SD,PREDICTED (blue), TREE_SD (orange), PREDICTED (yellow). Bottom: PLS error model using all metrics vs the RF model using all metrics.

TREE_SD,PREDICTED (blue) are very close to the diagonal. Models with a single descriptor TREE_SD (orange) and PREDICTED (yellow) are not as close. This seems to imply that one needs only TREE_SD and PREDICTED in combination to make a reasonable error model for our data sets. The usefulness of this will be addressed in the Discussion.

We tried PLS as an alternative to RF to make error models for SQRT(UE) based on its use in ref 22 and found that the R^2 for PLS is lower than for RF (Figure 7 bottom). The most obvious explanation is that PLS is not able to capture the nonlinear nature of the DA metric PREDICTED, which we have seen is an important descriptor.

Comparison of SQRT(UE) from Prospective Prediction. As with activity models, one would like to show that error models make reasonable predictions for the errors on molecules assayed after the model was built. Figure 8 shows scatterplots for the predicted and observed SQRT(UE) for three examples of prospective prediction: a good one (3A4), a medium one (PGP), and a poor one (HPLC_LOGD). Figure 9 (top) shows the R^2 for the predicted and observed SQRT(UE) for prospective prediction using all DA metrics vs the same using only the metrics TREE_SD and PREDICTED. As with the analogous situation for the cross-validated R^2 (the blue line in Figure 7), reducing the set of metrics to just those two does not significantly lower the accuracy of the prospective prediction of SQRT(UE). Figure 9 (bottom) compares the R^2 for prospective prediction for all DA metrics to the cross-validated R^2 for all DA metrics. Whereas in activity models one often sees that cross-validated R^2 is systematically higher than prospective R^2 (examples in ref 32), for the error models in this work the cross-validated R^2 is slightly lower.

Comparison of RMSE from Prospective Prediction. A plot of the observed RMSE vs the predicted RMSE for the prospective predictions is shown in Figure 10. Each circle

dx.doi.org/10.1021/ci400482e | J. Chem. Inf. Model. 2013, 53, 2837−2850

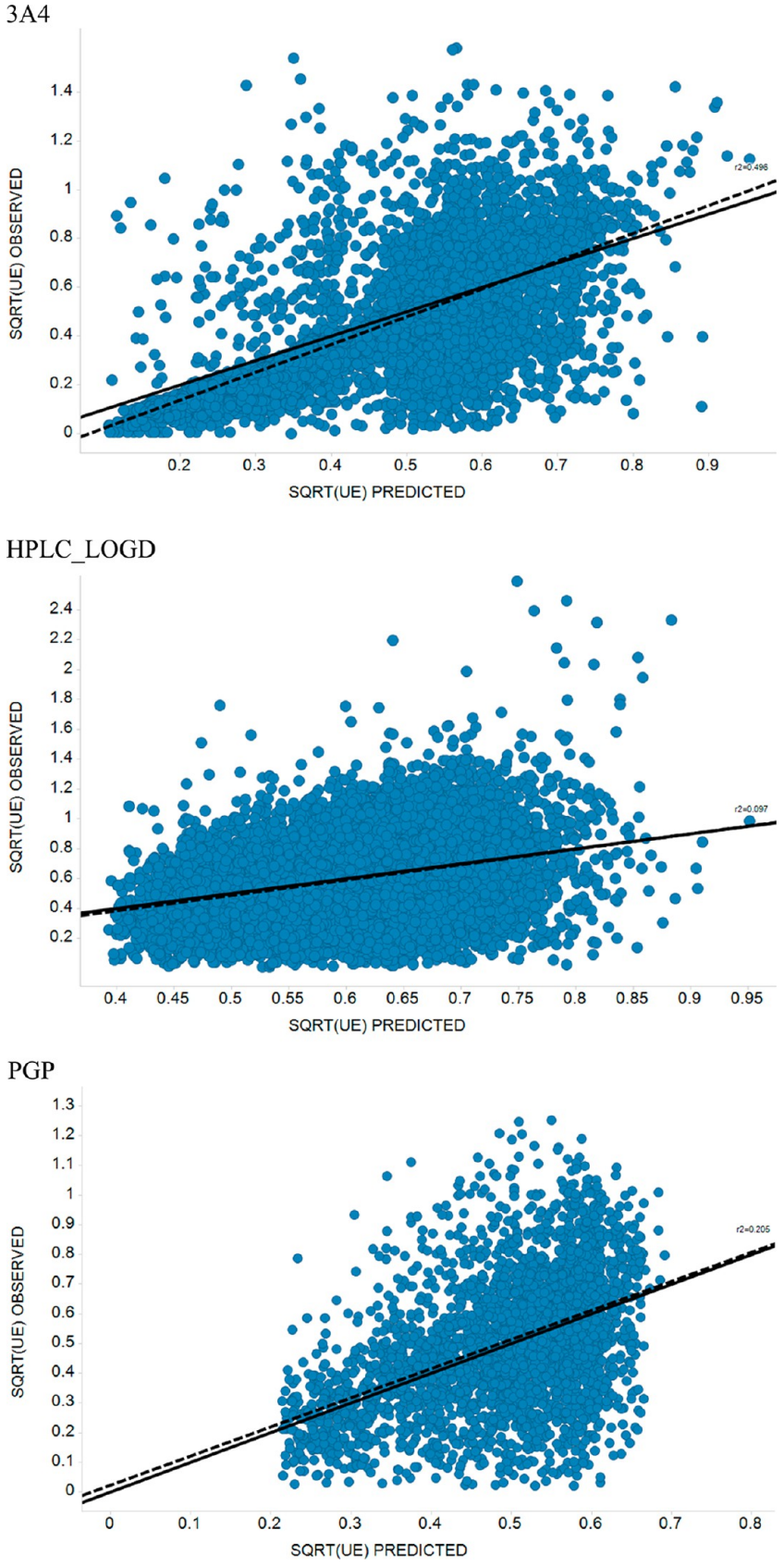


Figure 8. Relationship between the observed SQRT(UE) for the prospective prediction vs predicted SQRT(UE) from the RF error model using all metrics. The solid line is the diagonal, and the dotted line is the best linear fit.

encodes information about prospective prediction of one molecule M. Three types of error model are represented. The red points represent the RF error model using all metrics. The blue points represent the RF error model with only TREE_SD

dx.doi.org/10.1021/ci400482e | J. Chem. Inf. Model. 2013, 53, 2837−2850

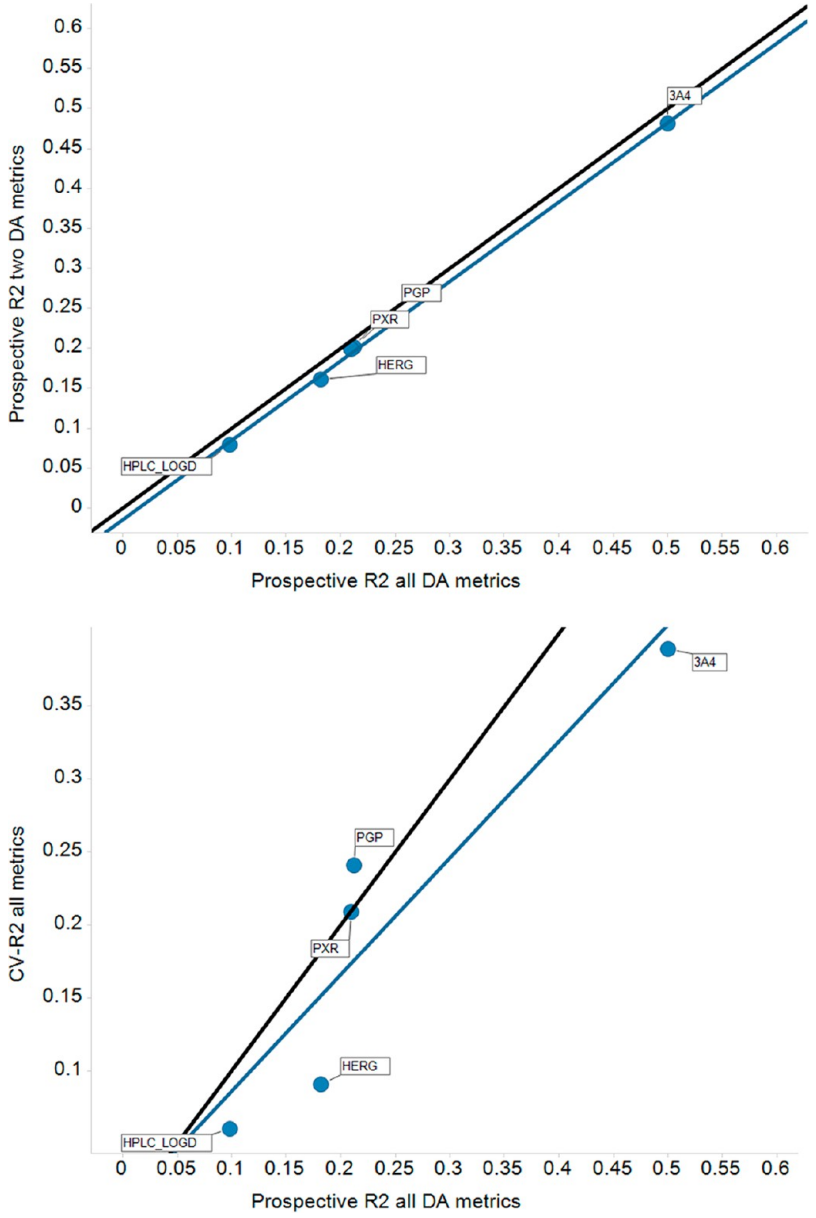


Figure 9. Top: R^2 for SQRT(UE) for prospective prediction using all DA metrics compared to using two metrics TREE_SD, PREDICTED. Bottom: R^2 for SQRT(UE) for prospective prediction using all DA metrics compared to the cross-validated R^2 using all DA metrics (bottom).

and PREDICTED as metrics. The green points represent 3DBINS. For all models in Figure 10 the relationship between predicted and observed RMSE is more or less linear, with the curves not far from the diagonal. The exception is HPLC_LOGD, not surprising because we know it has the poorest error model.

In Figure 10 the red and blue curves are nearly superimposed. This is expected because the all metric model and the model including only TREE_SD and PREDICTED give similar predicted RMSE. This is confirmed in the scatterplots in Figure 11. Again we see that HPLC_LOGD, the poorest model, does not show such a good agreement. 3DBINS is also fairly close to the others, not surprising since it also encodes information about the most important descriptors TREE_SD and PREDICTED, albeit in a different form. In about half the data sets we have examined, including some not in this paper, the 3DBINS system makes a slightly better absolute agreement between observed and predicted RMSE (is closer to the diagonal) than either RF model. Whatever the advantages of RF error models over the 3DBINS error model, better absolute prediction of prospective RMSE is not among them.

It should be noted that the apparent correlation in Figure 10 is very much higher than the apparent correlation in Figure 8. This is due to the phenomenon of "correlation inflation"[33] where a weak trend can appear stronger if one of the axes represents an average over many molecules. The y-coordinate in Figure 10 represents the RMSE calculated for many molecules, whereas the y-coordinate in Figure 8 represents the observed SQRT(UE) for a single molecule.

DISCUSSION

Whereas there is a general consensus of how well different QSAR methods work for building activity models, with variations on RF and SVM being the current favorites,[34,35] there is much less consensus how to build error models, which DA metrics are the most useful, whether they need to be used in combination, etc.

dx.doi.org/10.1021/ci400482e | J. Chem. Inf. Model. 2013, 53, 2837−2850

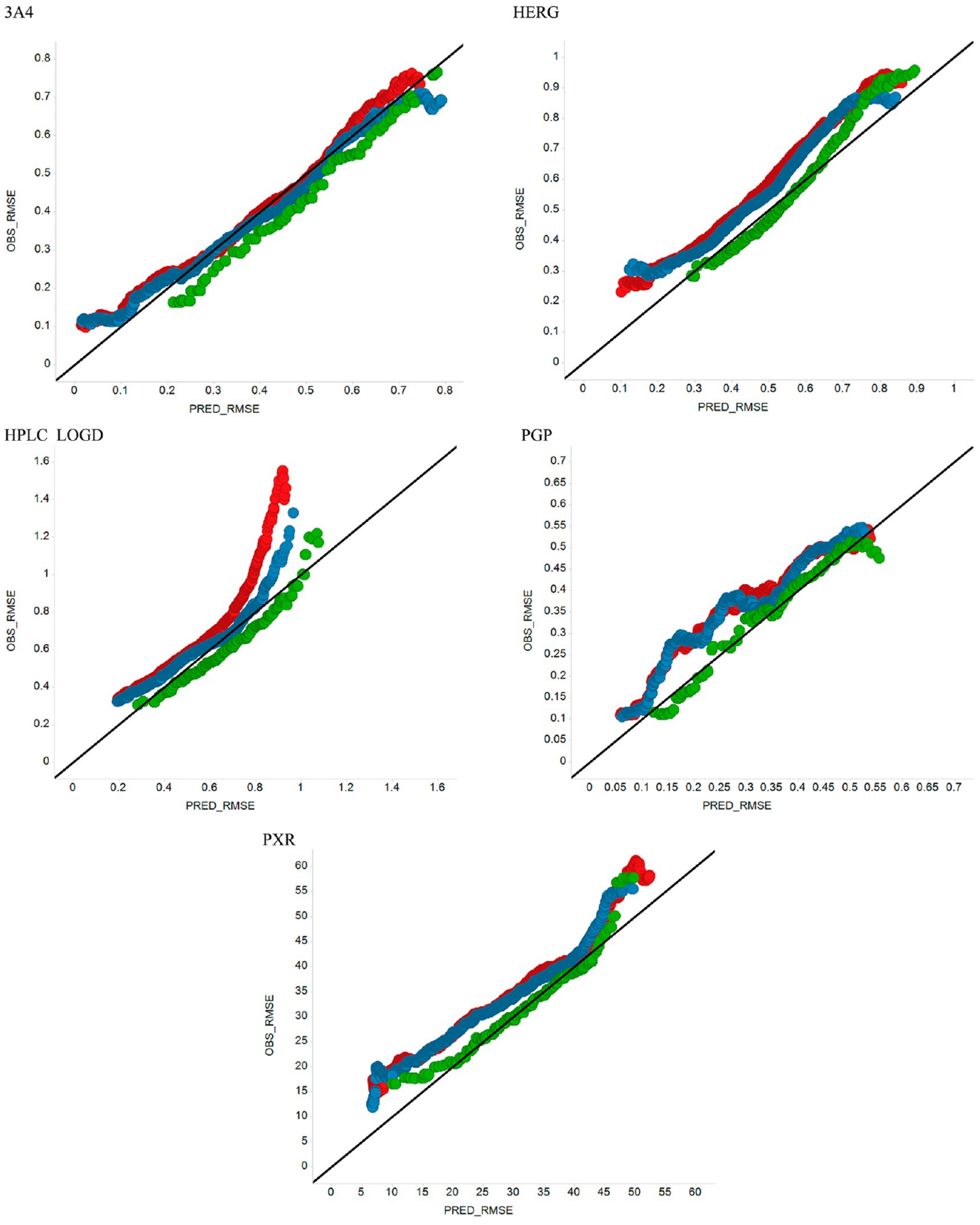


Figure 10. Observed RMSE of prospective prediction vs the RMSE predicted from the error models. The predicted RMSE is that of each molecule *M* being predicted. The observed RMSE is calculated for all compounds with a predicted RMSE within a small window around the predicted RMSE of M. Different colors represent different types of error models: RF model all metrics (red), RF model using only TREE_SD,PREDICTED (blue), 3DBINS model (green). Circles that correspond to sets of fewer than 100 molecules are not shown.

Each research group uses its own method for domain applicability. In the literature one sees a number of good visual trends of RMSE vs specific DA metrics. Our own early work[2] is no exception. In some cases, however, these trends may not be as

dx.doi.org/10.1021/ci400482e | J. Chem. Inf. Model. 2013, 53, 2837−2850

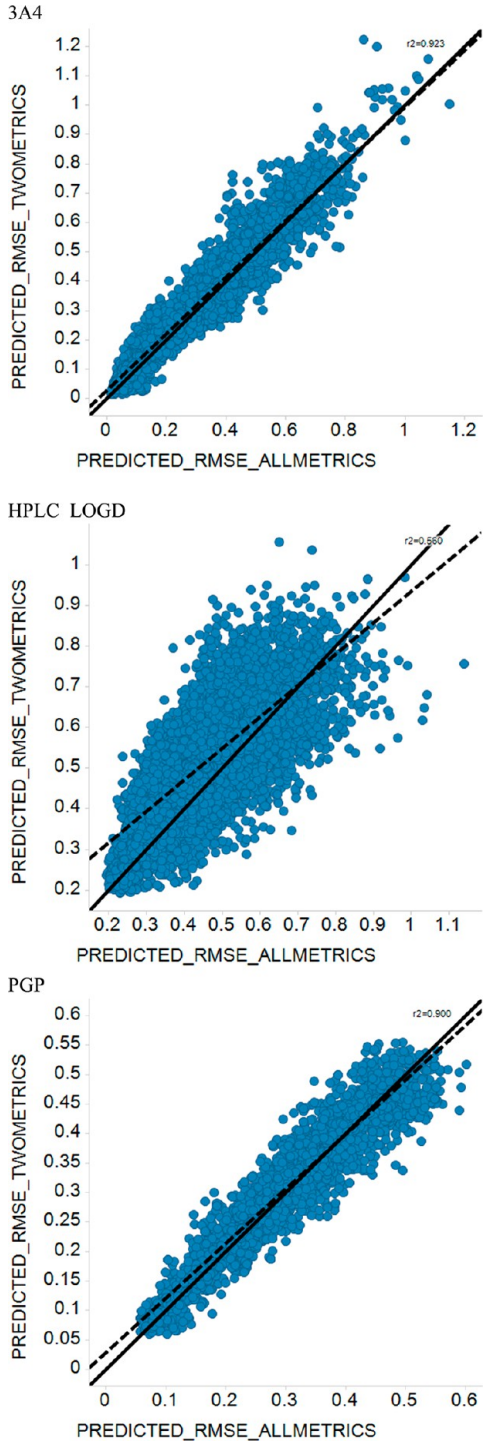


Figure 11. Relationship of predicted RMSE from the RF error model using TREE_SD, PREDICTED (TWO_METRICS) vs the predicted RMS from the RF model using all metrics. The black line represents the diagonal, and the dashed line represents the best linear fit.

strong as they appear. A trend that is actually very weak may appear strong if there is binning of the DA metric with many molecules per bin, which is necessarily the case with the calculation of RMSE. This phenomenon has been noted in the field of drug-likeness.[33] The application of QSAR methodology to building error models on more than one DA metric simultaneously gives us a chance to reveal such issues and to make comparisons of the usefulness of individual metrics because QSAR brings along standard practices such as validating models by cross-validation and defining descriptor importance. Here we use RF to generate the error models because it has a number of useful features. However, there are probably other QSAR methods that would be equally suitable, as long as they can handle nonlinear relationships.

It appears in retrospect that we may were misled in some of our earlier work with 3DBINS error models[18] the construction of which does not use the type of validation that we can use with QSAR-based error models. It was not obvious from inspection of the 3DBINS, for instance, that our HPLC_LOGD error model was not very good at distinguishing small from large errors. Also we thought that SIMILARITYNEAREST1 is more useful than we now believe. Another issue with our 3DBINS approach is that it required some arbitrary decisions on the part of the user. One critical decision is which three DA metrics one was going to use. Other decisions had to do with the bounds of the box in each dimension and the size of the bins. A number of compromises needed to be made. For instance, one could choose wide bins in which case one could accumulate a large number of predictions but limit the resolution of the model. Choosing narrow bins would have the opposite effect. Using a QSAR method like random forest relieves the user of making these decisions; the method automatically divides the DA metric space to best fit the error data.

One slight advantage of 3DBINS over the QSAR paradigm of model errors is that RMSE stored in the bins can be directly visualized in the space of three DA metrics. QSAR models like RF are generally not visualized directly. However, one can easily build a grid in any number of dimensions and predict the RMSE at each grid point. The grid can be easily visualized two or three dimensions at a time.

One philosophical concern about using a QSAR method for an error model is that one faces a potential "infinite regress". That is, if the activity model built with a QSAR method suffers from domain applicability issues, i.e., some predictions are more reliable than others, the error model would also suffer from the same issue. Ultimately we decided that this concern could apply to any type of model, not just those using QSAR methods. Domain applicability is probably less of an issue for error models than for activity models because the dimensionality of the DA metric space is small compared to the dimensionality of chemical descriptor space.

The original intent of using a QSAR method to model the errors of a QSAR activity model was to be able to properly handle a large number of DA metrics simultaneously. It is a small irony that, at least for large diverse data sets such as the ones we use here, it is necessary to include only two metrics TREE_SD and PREDICTED to get a reasonable estimate of the reliability of a prediction on the activity model. This is somewhat counter-intuitive. It is appealing to think that "similarity/distance to model" metrics should be important, and most early work in DA dealt with them. It was also somewhat of a surprise to us that the "local fit" metrics wRMSD1 and wRMS2 were ultimately less important than TREE_SD and PREDICTED in our examples. However, it has been recognized for a long time[9,11] that "bagged variance" DA metrics like TREE_SD seem to be very important, and more recent work[18,22] agrees. As far as we know, there is no other publication besides our own[18] that has tested PREDICTED as a metric. However, at least one other publication has noted that errors can depend on what part of the activity range the prediction is made.[1] That we need only TREE_SD and PREDICTED is useful because most of the other metrics require

one to calculate the similarity of each molecule *M* being predicted against all the molecules in the training set used to build the activity model. This is often a rate-limiting step with large data sets. We note however, that it is still possible that local data sets containing only close analogs may still require some "distance to training set" DA metric to correctly define the "domain" of the model.

Given how useful it is to model prediction errors with QSAR methods, it is surprising that few other workers have done this. We know of only two published cases. Guha and Jurs,[4] in one of the earliest papers on domain applicability, classified predictions on a linear model of boiling points as ″accurate″ and ″inaccurate″ and made a classification model of the errors using a neural network. Interesting, the ″descriptors″ used with the neural network error model were the original chemical descriptors used with the activity model and not DA metrics. A very recent paper by Wood et al.[22] constructed an error model of UE from cross-validation vs three DA metrics using PLS. Wood et al. also anticipated some of our findings. They found that the ″bagged variance″ DA metric from RF by itself provides a reasonable error model. They also noted for their LOGD example that even though one might be able to accurately predict an activity, one might not be able to accurately predict the error bars on that activity.

■ ASSOCIATED CONTENT

Supporting Information

Activity data set, including SMILES and descriptors for the FACTORX example, and error data set for FACTORX. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: sheridan@merck.com.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

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. *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) Gua, R.; Van Drie, J. H. Structure-activity landscape index: quantifying activity cliffs. *J. Chem. Inf. Model.* **2008**, *43*, 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.; Öerg, 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 method 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*, 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.; Stalring, 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) 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.

(24) 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.
(25) 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.
(26) PubChem. http://pubchem.ncbi.nlm.nih.gov/ (accessed Oct. 1, 2011).
(27) Sun, H.; Veith, H.; Xia, M.; Austin, C. P.; Huang, R. Predictive models for cytochrome P450 isozymes based on quantitative high throughput screening data. *J. Chem. Inf. Model.* **2011**, *51*, 2474−2481.
(28) National Center for Biotechnology Information. PubChem BioAssay Database; AID=1851, Source=Scripps Research Institute Molecular Screening Center. http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=1851 (accessed Oct. 8, 2013).
(29) http://dtp.nci.nih.gov/docs/aids/aids_data.html (accessed Oct. 1, 2011).
(30) *ChEMBL*. https://www.ebi.ac.uk/chembl/ (accessed February 14, 2012).
(31) National Center for Biotechnology Information. PubChem BioAssay Database; AID=361, Source=Scripps Research Institute Molecular Screening Center. http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=361 (accessed Oct. 1, 2011).
(32) Sheridan, R. P. Time-split cross-validation as a method for estimating the goodness of prospective prediction. *J. Chem. Inf. Model* **2013**, *53*, 783−790.
(33) Kenny, P. W.; Montanari, C. A. Inflation of correlation in the pursuit of drug-likeness. J. *Comput.-Aided Mol. Des.* **2013**, *27*, 1−13.
(34) Bruce, C. L.; Melville, J. L.; Picket, S. D.; Hirst, J. D. Contemporary QSAR classifiers compared. *J. Chem. Inf. Model.* **2007**, *47*, 219−227.
(35) Svetnik, V.; Wang, T.; Tong, C.; Liaw, A.; Sheridan, R. P.; Song, Q. Boosting: an ensemble learning tool for compound classification and QSAR modeling. *J. Chem. Inf Model.* **2005**, *45*, 786−799.