JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

Interpretable, Probability-Based Confidence Metric for Continuous Quantitative Structure−Activity Relationship Models

Christopher E. Keefer,*[,†] Gregory W. Kauffman,[‡] and Rishi Raj Gupta[§,⊥]

[†]Computational ADME Group, Department of Pharmacokinetics, Dynamics, and Drug Metabolism, and [§]Research Centers of Emphasis, Pfizer Inc., Groton, Connecticut 06340, United States

[‡]Worldwide Medicinal Chemistry, Neuroscience Research Unit, Pfizer Inc., Cambridge, Massachusetts 02139, United States

Supporting Information

ABSTRACT: A great deal of research has gone into the development of robust confidence in prediction and applicability domain (AD) measures for quantitative structure−activity relationship (QSAR) models in recent years. Much of the attention has historically focused on structural similarity, which can be defined in many forms and flavors. A concept that is frequently overlooked in the realm of the QSAR applicability domain is how the local activity landscape plays a role in how accurate a prediction is or is not. In this work, we describe an approach that pairs information about both the chemical similarity and activity landscape of a test compound's neighborhood into a single calculated confidence value. We also present an approach for converting this value into an interpretable confidence metric that has a simple and informative meaning across data sets. The approach will be introduced to the reader in the context of models built upon four diverse literature data sets. The steps we will outline include the definition of similarity used to determine nearest neighbors (NN), how we incorporate the NN activity landscape with a similarity-weighted root-mean-square distance (wRMSD) value, and how that value is then calibrated to generate an intuitive confidence metric for prospective application. Finally, we will illustrate the prospective performance of the approach on five proprietary models whose predictions and confidence metrics have been tracked for more than a year.

INTRODUCTION

Quantitative structure−activity relationships (QSARs) have enjoyed an extensive vetting in the literature since their inception in the 1960s.[1] A bounty of studies reporting "successful" QSAR models have been reported; however, the methodology does have its skeptics and many skilled QSAR modelers have offered critiques[2−6] and guidance over the past few years.[7−11] One of the key criticisms of QSAR models is their general inability to predict the activity of new compounds accurately and the uncertainty of knowing when that is the case. Assuming a QSAR model has been built and validated to the highest standard, inaccurate predictions can generally follow for two reasons: (1) the new compound is not similar enough to the chemistry space of the model training set and/or (2) the activity space of the model is fraught with activity cliffs that are not captured by the model. Neither of these two cases necessarily renders the underlying QSAR model useless, but they do highlight that the end-user needs, in addition to the prediction itself, some indication that the prediction should or should not be trusted. To this end, a variety of confidence-in-prediction and applicability domain (AD) measures have been introduced in the QSAR field over the past decade.[12−18]

In the mid-2000s, the Pfizer design community largely transitioned from classification-based in silico statistical models to regression-based models. One of the inherent benefits of classification models is that one can simply use the ratio of correct bin assignments across a committee of bagged or boosted models to arrive at a "confidence" in each prediction. Regression-based models do not lend themselves to such a simple, intuitive representation of confidence, which the Pfizer design community desired under the new modeling paradigm. At the time, structural nearest neighbor approaches were receiving much attention for this purpose in the literature.[19] To meet the needs of design teams, a simple "number of nearest neighbors" approach was implemented to assign prediction confidence for new target compounds. In our implementation, a nearest neighbor was defined as a training set compound having a pairwise atom-pair fingerprint Tanimoto coefficient above a certain threshold. The coefficient threshold varied among end points and modeling methods, but typical values were between 0.7 and 0.8. The naïve assumption with this approach was that a proposed target compound with a greater number of nearest neighbors in the model training set was more likely to have a lower prediction error than a compound with fewer neighbors. While this assumption held largely true, in reality throngs of compounds with few nearest neighbors had low prediction errors. In addition, compounds at the bounds of

Received: November 19, 2012
Published: January 23, 2013

Table 1. Parameters for Literature Data Sets

name	end point	ref	*N*; prefilter	*N*; postfilter	model descriptors (dragon block)	error threshold
AIDS	$-\log_{10}(EC_{50})$	23	36396	34002	topological indices (3) atom-centered fragments (22)	2-fold
CYP1A2	$-\log_{10}(AC_{50})$	24	17143	15811	ring descriptors (2) atom-centered fragments (22)	2-fold
Vd_{ss}	$\log_{10}(Vd_{ss}$ (L/kg))	25	670	592	2D matrix-based descriptors (7) molecular properties (28)	2-fold
DHFR	$-\log_{10}(IC_{50})$	26	756	720	topological indices (3) CATS 2D (24)	2-fold

activity cliffs could also have many neighbors but still be poorly predicted. These observations led us to conclude that structural/chemical similarity alone was insufficient to establish confidence in a prediction.

In 2008, an in-house re-evaluation of prediction confidence metrics was undertaken to improve upon the nearest neighbor methodology. The overarching goal of this effort was to develop an intuitive confidence metric, preferably in the range of 0.0−1.0, that was modeling method independent and that augmented the chemical space of nearest neighbors with information about the *activity space* of those neighbors as well. A survey of the literature revealed, at the time, that although there are QSAR methods that account for both structural and activity space similarity in their predictions,[20,21] AD methods focused almost exclusively on the chemical similarity component, with no meaningful effort to address the activity space component. We found this intriguing, as nearly every data set of bioactive end points will invariably contain activity cliffs, and we believed that incorporating information about the structure−activity landscape of the nearest neighbor space would enhance the quality of any prediction confidence metric. Recently, Sheridan described a three-parameter, random forest-specific AD approach that includes as one of its terms the predicted value of a compound.[12] His justification for including this parameter is the presumption that different regions along the activity range will model better than others, which those skilled in the art of QSAR know to be invariably true.

This paper will describe a new confidence metric for in silico activity modeling that incorporates both the chemical similarity and the activity landscape of the nearest neighbors for a test compound. The implementation of the method will be described and illustrated for four literature data sets of varying size and composition. The practical utility and prospective performance of the metric will then be demonstrated with five proprietary data sets and models that are actively used in design projects at Pfizer. This will be followed by an in-depth discussion about the pros and cons of the method, plus some general observations about the behavior of this metric in a real-world setting.

METHODS

Data Sets. The implementation and underlying assumptions, strengths, and weaknesses of this confidence metric will be demonstrated with models built on four data sets from the literature: AIDS antiviral screen (AIDS), CYP1A2 inhibition, human volume of distribution at steady state (Vd_{ss}), and dihydrofolate reductase inhibition (DHFR). This will be followed by an analysis of the prospective performance of the metric, over a 1−2 year period, for in silico models built on five proprietary Pfizer data sets. The data sets include four high-throughput end points[22] (human liver microsome (HLM) metabolic stability, shake-flask logD, P-gp efflux ratio, RRCK membrane permeability) and one project-specific GCPR target binding assay.

Prior to modeling, both the literature and in-house data sets were filtered to remove end point values that were unreasonably high or low, chemical structures that contain atom types other than C, N, O, P, S, Si, B, F, Cl, Br, I, or H, and structures that contain supervalent atoms or zero carbon content. The filtered literature data sets were then split into training, test, and validation sets using a venetian blind selection process. In short, this process is done as follows. First, the data are sorted by activity and repeatedly numbered 1−5. Then, all the data labeled 1, 2, or 3 is pulled as the training set, the data labeled 4 is selected as the test set, and the data labeled 5 is selected as the external validation set. This results in a 60% training, 20% test, and 20% validation split of the data. One advantage of a sorted venetian blind selection is that it maintains similar activity distributions in each of the derived sets even when the overall data is heavily skewed as is often the case in large compound activity data sets. Finally, the activity end points for each data set were transformed to an appropriate log scale for modeling purposes. Data set source, sizes, both pre- and postfiltering counts, as well as end point transformations for each model are given in Table 1.

Descriptors. For the four models built from literature data sets, descriptors were computed with the Dragon software package.[27] The Dragon descriptors are organized into 29 blocks according to their information content. To find suitable descriptors for modeling each data set, the performance of all two-block combinations of 2D descriptors were evaluated using the Cubist[28] modeling method, which is later described in the Modeling Methods section. The descriptor combination resulting in the highest Pearson correlation coefficient (R) on the hold-out test set was selected as the final model. The descriptor blocks used for each model are listed in Table 1. Note that this was not an exhaustive search for optimal subsets of descriptors within Dragon blocks, as the goal of this work was to build models of sufficient quality to demonstrate the performance of the confidence metric.

The descriptors used for the proprietary models are more varied. The four high-throughput ADME models were built using optimized combinations of Dragon descriptor blocks, logP, and logD from the MoKa Suite[29] and a collection of proprietary, substructural fragment counts tailored for modeling ADME/T end points. The GPCR activity model employs 2D descriptors from the MOE software package[30] and proprietary substructural fragment counts. All in-house models were built using Cubist.

Modeling Methods. To demonstrate the performance of the confidence metric across different modeling methods, models were built for the literature data sets using three

dx.doi.org/10.1021/ci300554t | J. Chem. Inf. Model. 2013, 53, 368−383

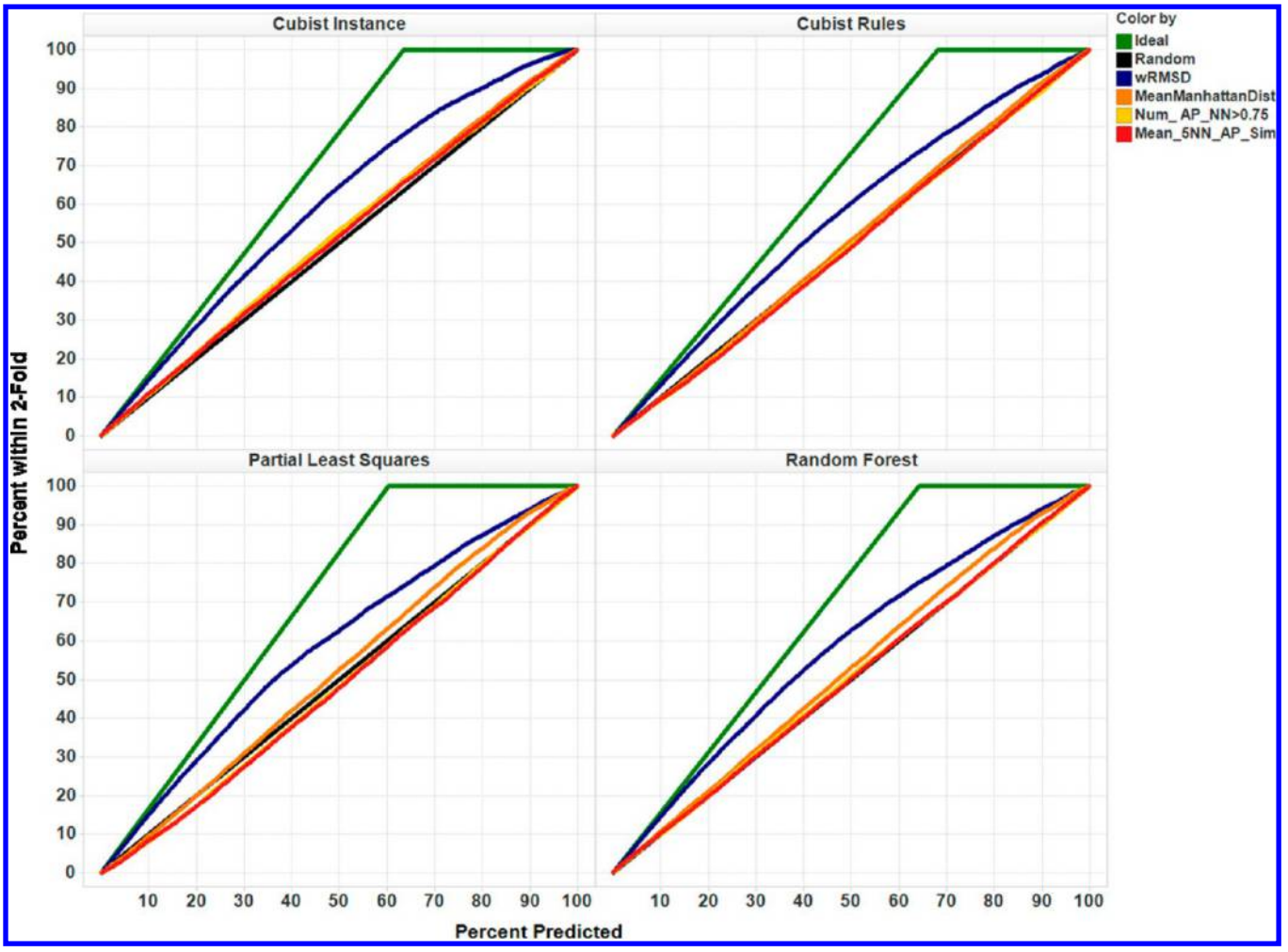

Figure 1. Chart comparing wRMSD to chemical distance/similarity methods for the AIDS data set.

regression techniques: Cubist, partial least squares (PLS), and random forest (RF). Models were built on the training set compounds and the model parameter selection for PLS, and RF was based on cross-validation prediction performance. In addition, the confidence value calibration step, described in more detail later, is done using the test set. Final performance of the model and confidence metric was evaluated with the validation set, which was completely blinded from the model building and confidence value calibration processes. All proprietary models discussed in this paper were built with Cubist using the same validation and calibration approaches.

Cubist is recursive partitioning based modeling software from RuleQuest Research.[21,28,31] There is a free GPL licensed Linux source code version as well as a commercial multithreaded version for Windows and Linux. Cubist builds two types of models: rule-based models and instance models (also called composite models). In an instance model, the rule-based prediction of a compound is "corrected" using the prediction errors of its training set nearest neighbors. In this work, results for both Cubist model types are reported. Models were built using 20 committees, allowing up to 1000 rules per committee. For the instance models, the maximum nearest neighbors parameter was set to five. The PLS and random forest models were built with the Caret package[32] in the R statistical computing environment (version 2.15.0)[33] using 10-fold cross-validation and default parameters. The R script used to build the models and make test and training set predictions is included in the Supporting Information.

Calculating Chemical Similarity and Defining Nearest Neighbors. A key element of the confidence metric described in this work is the appropriate selection of training set nearest neighbors for a test compound. There are myriad ways to define structural nearest neighbors. One effective approach, and the method employed in this work, is the modified Manhattan (Hamming) distance used in Cubist to identify nearest neighbors for composite models.[34] The Cubist distance between any two compounds is defined by eq 1

$$D = \sum_{i=1}^{p} \min\left(\frac{|x_{1i} - x_{2i}|}{5\sigma_i}, 1.0\right) \quad (1)$$

where p is the total number of descriptors in the data set, x_{1i} is the descriptor value for the first compound's ith descriptor, x_{2i} is the descriptor value for the second compound's ith descriptor, and σ_i is the standard deviation of the ith descriptor for the full training set. This equation normalizes each of the descriptor distances by its standard deviation and sets the maximum range of distances within a descriptor to 0 to 5 standard deviations (0.0−1.0 in D). The minimum function sets any distance greater than 5 standard deviations to 1.0. For purposes of selecting training set nearest neighbors for a new test compound, those neighbors with the lowest values of D

dx.doi.org/10.1021/ci300554t | J. Chem. Inf. Model. 2013, 53, 368−383

should define the most relevant chemical space in the training set.

Calculating Weighted RMSD. As mentioned previously, structural/chemical similarity is only part of the formula for defining prediction confidence. To incorporate the activity space similarity of a new test compound to its training set NNs, a variety of formulas that incorporate activity and distance were investigated. Ultimately, the best performance was achieved with the similarity-weighted root-mean-squared distance given by eq 2

$$\text{wRMSD} = \sqrt{\frac{\sum_{i=1}^{N} w_i^2(\hat{y} - y_i)^2}{\sum_{i=1}^{N} w_i^2}} \quad (2)$$

where N is the number of nearest neighbors used in the confidence calculation, $\hat{y}$ is the predicted response of the compound, y_i is the experimental activity of the compound's ith neighbor, and w_i is the similarity weight of the ith neighbor, given by eq 3

$$w_i = \frac{1}{D_i + 0.5} \quad (3)$$

where D_i is the distance to the ith neighbor as defined in eq 1. In eq 3, 0.5 is added to D_i in the denominator to prevent division by zero errors when the distance is 0.0.

In this work, an N of five was used for all models and confidence calculations. The wRMSD is a "smaller is better" value meaning lower values approaching 0.0 are indicative of high confidence. What this conceptually means is that when wRMSD is low, the predicted activity for a new test case is consistent with the experimental activity of its five nearest neighbors, and the activity among those neighbors is similar representing a smooth SAR landscape. On the other hand, higher values of wRMSD indicate low confidence, which can occur for two reasons. The first scenario is when the predicted activity for a new test case is dissimilar from the experimental values of its nearest neighbors, but the consistency among the neighbors is good. Clearly, some feature of the new compound is at odds with its neighbors and presents itself as a potential activity cliff. The second scenario is when the activity space of the nearest neighbors varies (local activity cliff SAR landscape), regardless of whether the predicted activity of the new test compound is similar or dissimilar to some of the neighbors. In this case, there is too much uncertainty surrounding the SAR of the neighbors for the user to be certain of the model prediction on the new test compound.

Comparison of wRMSD to Distance Based Metrics. To compare the performance of the calculated wRMSD method to traditional applicability domain methods that utilize chemistry space similarity alone, we examined lift charts to assess the ability of each method to rank order predictions that are within 2-fold of their experimental values for the validation set. Figure 1 shows lift charts for the AIDS data set and the four modeling methods. The x-axis represents the percentage of compounds in the data set (the top x% of compounds by rank) and the y-axis represents the percentage of total compounds whose prediction is within 2-fold of its experimental value. The green line represents the ideal case where all of the compounds within 2-fold are ranked at the top of the list, and the black line represents the random case where there is no enrichment and the rate of discovery of compounds within 2-fold is the same as the overall rate in the data set. The wRMSD value is represented by the blue line and the mean Manhattan distance for the 5 nearest neighbors in the training set, using the same descriptors as those in the model, is represented by the orange line. Finally, there are two atom pair fingerprint based measures of similarity: the count of neighbors with a tanimoto similarity >0.75 shown in yellow and the mean tanimoto similarity of the five nearest neighbors in the training set shown in red.

The graph illustrates a significant enrichment using wRMSD versus the distance and similarity-based methods, supporting the assertion that both activity and chemical space are critical components for confidence assessment. The lift charts for CYP1A2 and DHFR (Supporting Information) illustrate a similar trend between wRMSD and the other three methods. For the Vd_{ss} data set, the performance of wRMSD is less robust but consistent with the other methods. This result is not completely unexpected as this data set is extremely diverse with few compounds from similar chemistry spaces. The lift charts for CYP1A2, DHFR, and Vd_{ss} are included in the Supporting Information.

Calibration of wRMSD. One limitation of the wRMSD value is that interpretation of the raw values is neither simple, nor straightforward. The challenge this presents is that the range of wRMSD values may vary from model to model, making it impossible to know a priori what range of values constitute high and low confidence in practice. To address the issue of interpretability, we developed a process that calibrates the raw wRMSD values with the accuracy of the test set predictions. The first step in this process is to establish a threshold that constitutes an acceptable level of prediction error. To do this, one must have a reasonable understanding of the variability in the data being modeled so that predictions can be deemed "good" or "bad" based on a threshold. In our process, thresholds can be set as a fold error or residual error (i.e., a good prediction is within 2-fold of experiment or has a residual error less than 0.5 units). In the second step, the predictions are sorted by wRMSD and the proportion of good predictions within a particular range of wRMSD is calculated. Two methods to define the range can be used, which are illustrated in Figure 2. The first method is a moving proportion over an equal width window, which we call the range moving proportion (RMP). Having examined several range sizes, we have found that a range of ±0.05 wRMSD units is suitable in

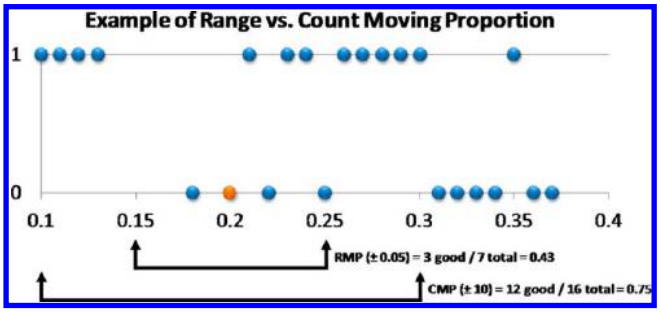

Figure 2. Example of RMP and CMP calculations. For a given point (0.2 in this case), the range moving proportion is the proportion of all the compounds within a specified range (±0.05) of the point in question that are good (value of 1 in this example). For the count moving proportion, the proportion is calculated over all of the compounds within a specified count of the point in question (±10) and includes the point in question. Note that, if there are not enough points due to range limitation, the proportions are calculated over the points that are present.

dx.doi.org/10.1021/ci300554t | J. Chem. Inf. Model. 2013, 53, 368−383

Table 2. Model Performance Statistics[a]

			Cubist instance			Cubist rules			partial least squares			random forest		
data set	model set	*N*	% good[a]	RMSE	*R*	% good[a]	RMSE	*R*	% good[a]	RMSE	*R*	% good[a]	RMSE	*R*
AIDS	validation	6801	62.7	0.47	0.54	66.1	0.51	0.45	58.1	0.51	0.38	62.6	0.47	0.55
CYP1A2	validation	3162	46.7	0.64	0.64	46.1	0.63	0.64	39.4	0.66	0.59	43.3	0.62	0.65
DHFR	validation	144	38.2	0.73	0.78	29.2	0.75	0.75	33.3	0.96	0.56	41.0	0.73	0.77
Vd_{ss}	validation	119	55.5	0.45	0.70	54.6	0.43	0.72	56.3	0.53	0.53	56.3	0.47	0.66
HLM	prospective	133189	63.3	0.36	0.71									
P-gp	prospective	102164	72.9	0.31	0.70									
SFLogD	prospective	66115	71.5	0.56	0.90									
RRCK	prospective	138016	67.7	0.40	0.75									
GPCR	prospective	1494	78.0	0.77	0.51									

[a]For the in-house, prospective data sets, only the Cubist instance results are shown. [b]% good is defined as the percentage of predictions within 2-fold of the actual value for all end points other than SFLogD where it is defined as the percentage of predictions within ±0.5 units.

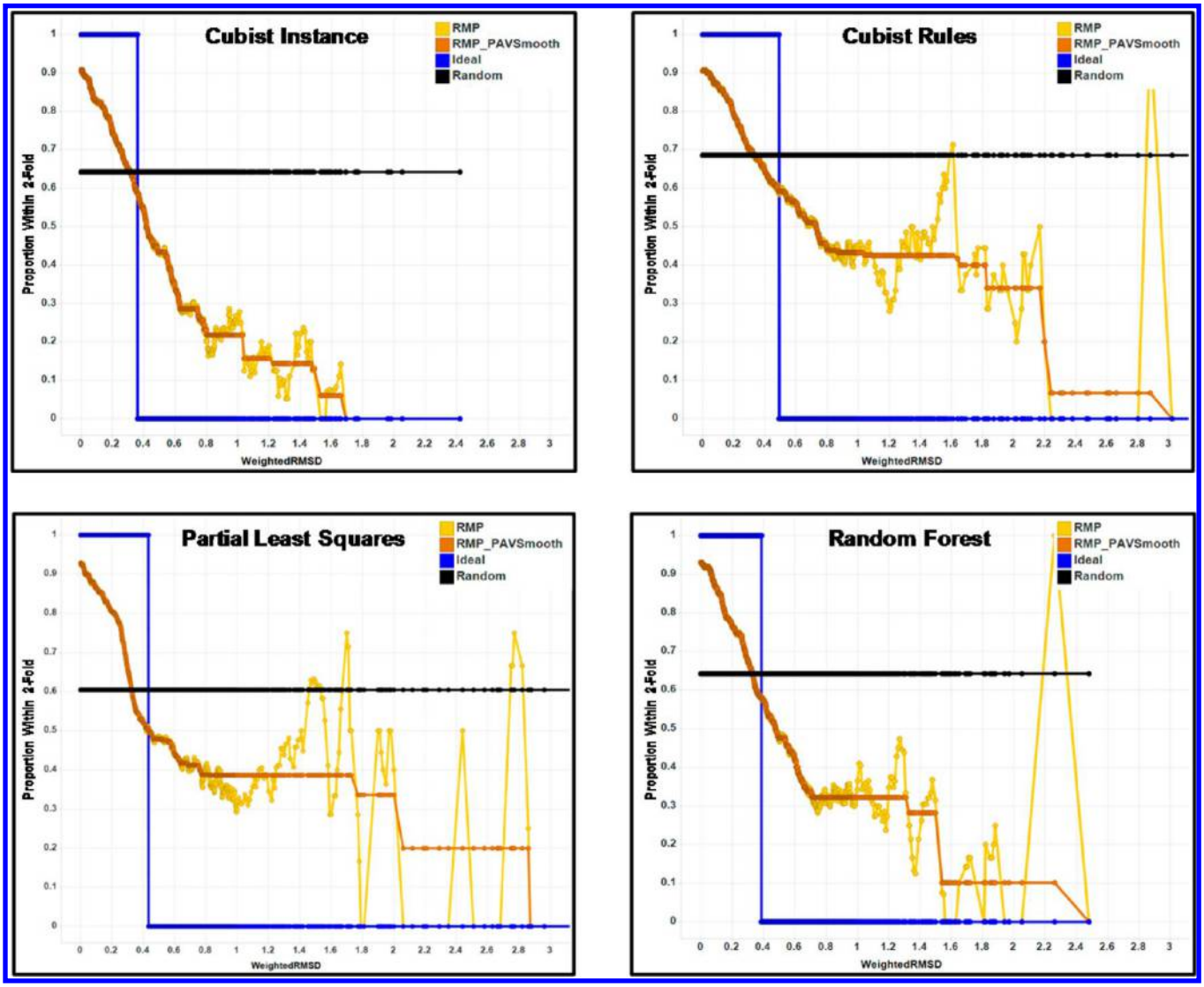

Figure 3. Calibration plots (2-fold) for the AIDS data set. The black line represents the overall proportion of predictions within 2-fold which is what would be expected if there was no improvement from the confidence metric. The blue line represents the ideal performance for a confidence metric; the case where all of the compounds within 2-fold have the highest confidence. The yellow line is the range moving proportion (RMP) of predictions within 2-fold. The orange line is the monotonic pool adjacent violators (PAV) smooth of the yellow line and is the fit used to generate the final confidence metric values.

most cases. In the example shown in Figure 2, the prediction for a theoretical test case has an accompanying wRMSD value of 0.2. To calculate the RMP, only the good predictions within the wRMSD range of 0.15−0.25 are considered. In this example, a total of seven predictions are within this range with three good predictions, which is a proportion of 43%. The

dx.doi.org/10.1021/ci300554t | J. Chem. Inf. Model. 2013, 53, 368−383

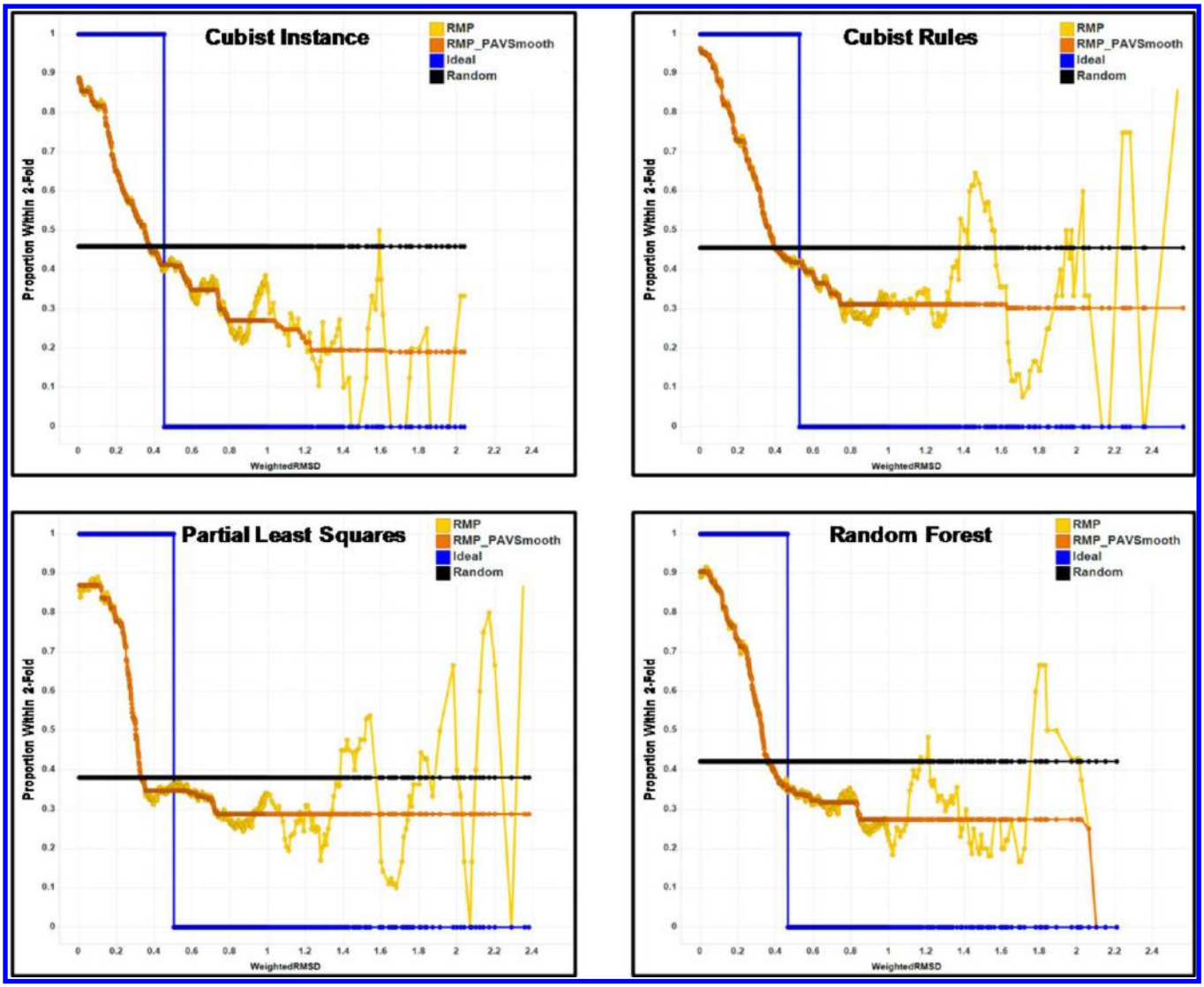

Figure 4. Calibration plots (2-fold) for the CYP1A2 data set. The black line represents the overall proportion of predictions within 2-fold which is what would be expected if there was no improvement from the confidence metric. The blue line represents the ideal performance for a confidence metric; the case where all of the compounds within 2-fold have the highest confidence. The yellow line is the range moving proportion (RMP) of predictions within 2-fold. The orange line is the monotonic pool adjacent violators (PAV) smooth of the yellow line and is the fit used to generate the final confidence metric values.

second method is a moving proportion over a fixed number of samples before and after the value of interest, which we call the count moving proportion (CMP). We have found that a range of ±10 observations is typically suitable for this approach, thus the moving proportion is calculated over a total of 21 data points. Using the same example from Figure 2 with a starting wRMSD value of 0.2, the CMP method moves up and down the range of wRMSD values until ±10 observations are captured or, in this case, will stop once the minimum or maximum value of wRMSD is reached. For the example, a total of 16 values are included using the CMP approach, resulting in a proportion of 75% good predictions. We have found that the CMP method works better for models with small, diverse test sets since there are often not enough observations in the equal width windows to calculate robust proportions with the RMP method. Therefore, we used the CMP method for the DHFR and Vd_{ss} models, and the RMP approach for the AIDS, CYP1A2, and all five proprietary models. For each of the test set predictions, the wRMSD and moving proportions are calculated and stored for the next step of the smoothing process.

In the final step of the calibration process, the pool adjacent violators (PAV) algorithm[35,36] is used to monotonically smooth the moving proportion values. This is done on the premise that as values of wRMSD increase, the confidence in the corresponding predictions should decrease. The monotonic smoothing starts with the first sample in the data set (the point with the lowest wRMSD value) $i = 1$. Next, the pair (P_i, P_{i+1}), where P is the moving proportion value calculated above, is checked for violations of the monotonicity constraint (i.e., check that $P_i > P_{i+1}$). If the constraint is not violated, increment i and check the next pair until all pairs have been checked. If the constraint is violated, P_i and P_{i+1} are replaced with the average of P_i and P_{i+1}. Since the value of P_i changed, a back check is performed to ensure that P_{i-1} is still greater than or equal to P_i. If this is not the case, P_{i-1}, P_i, and P_{i+1} are replaced with the average of all three of their values. This back check is repeated with prior values of P, and with all values being replaced with

dx.doi.org/10.1021/ci300554t | J. Chem. Inf. Model. 2013, 53, 368−383

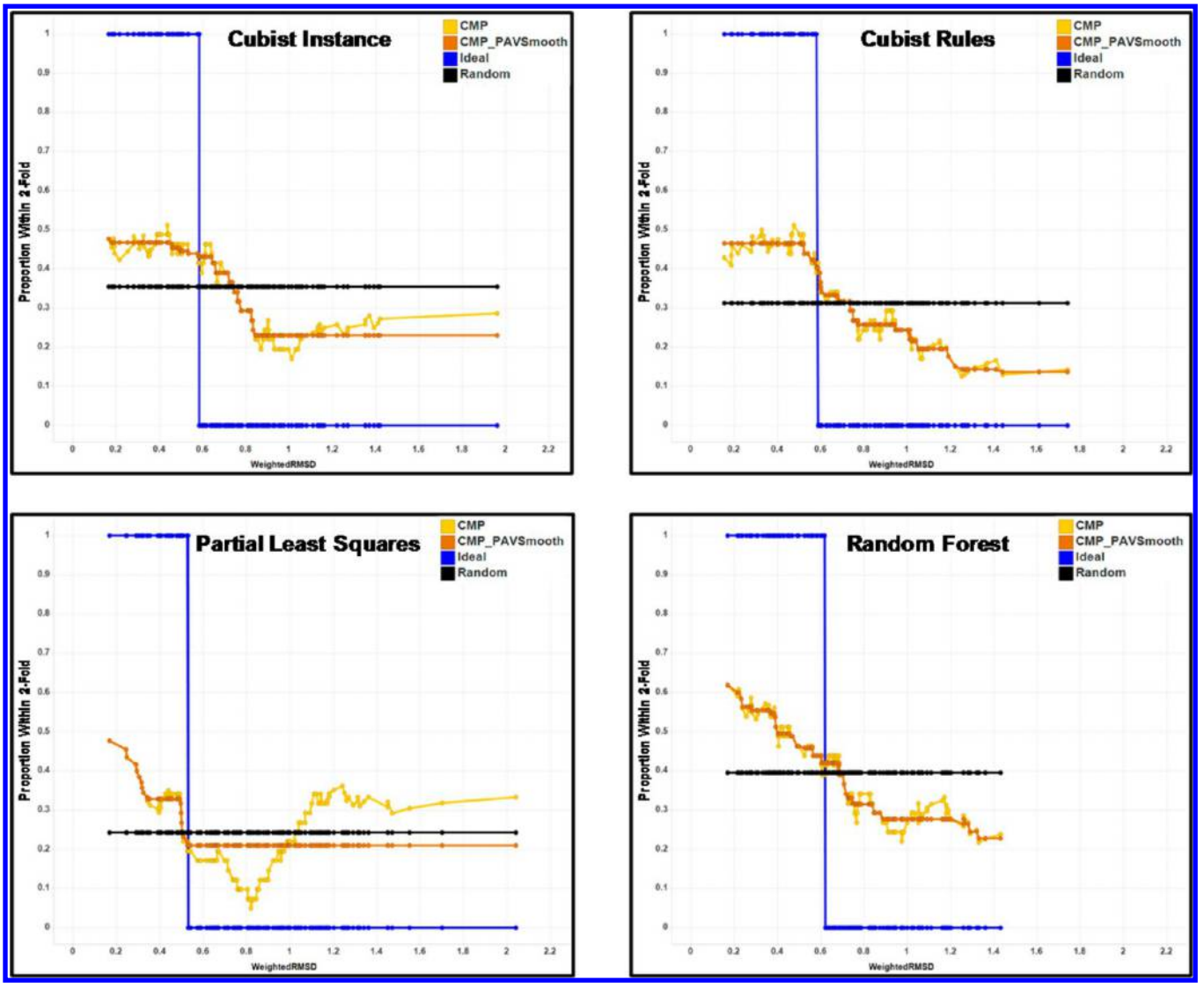

Figure 5. Calibration plots (2-fold) for the DHFR data set. The black line represents the overall proportion of predictions within 2-fold which is what would be expected if there was no improvement from the confidence metric. The blue line represents the ideal performance for a confidence metric; the case where all of the compounds within 2-fold have the highest confidence. The yellow line is the count moving proportion (CMP; n = 20) of predictions within 2-fold. The orange line is the monotonic pool adjacent violators (PAV) smooth of the yellow line and is the fit used to generate the final confidence metric values.

their average, until the monotonicity constraint is satisfied. Once the back checking constraint is satisfied, increment i and continue the pairwise checks until all compounds have been processed.

This monotonic smoothing step produces a table relating raw wRMSD values to a proportion of test compounds that are "good". These values can then be used as an estimated probability that a prediction is "good" given its wRMSD value. If the wRMSD value explicitly exists in the table, the smoothed proportion of "good" compounds is returned. If the wRMSD value does not exist in the table, the smoothed proportion is linearly interpolated from the two adjacent wRMSD values and their respective smoothed proportions. These interpolated probabilities of a prediction being good are the final confidence metric reported to users of the models.

Prospective Prediction Database. We are fortunate to have a variety of in silico models built upon data from proprietary, high-throughput ADMET and active therapeutic area projects to demonstrate the real-world utility and performance of the new confidence metric described herein. We also have the benefit of an extensive database of *prospective* model predictions and confidence values for a majority of our in-house models. To generate this database, model predictions and confidence values are captured at the time of compound registration. The values in this database remain static even after the models have been updated. Once the compounds have been tested experimentally, the stored predictions can be compared to the experimental data providing a truly prospective validation of the model.

RESULTS

Model Performance. In Table 2, the root-mean-squared error (RMSE), Pearson correlation coefficient (R), and percentage of "good" predictions are shown for all data sets used in this study. For the four literature sets, statistics for the test and validation sets are included. For the five proprietary data sets, only statistics for the prospective predictions are included. Overall model performance is similar among the four

dx.doi.org/10.1021/ci300554t | J. Chem. Inf. Model. 2013, 53, 368−383

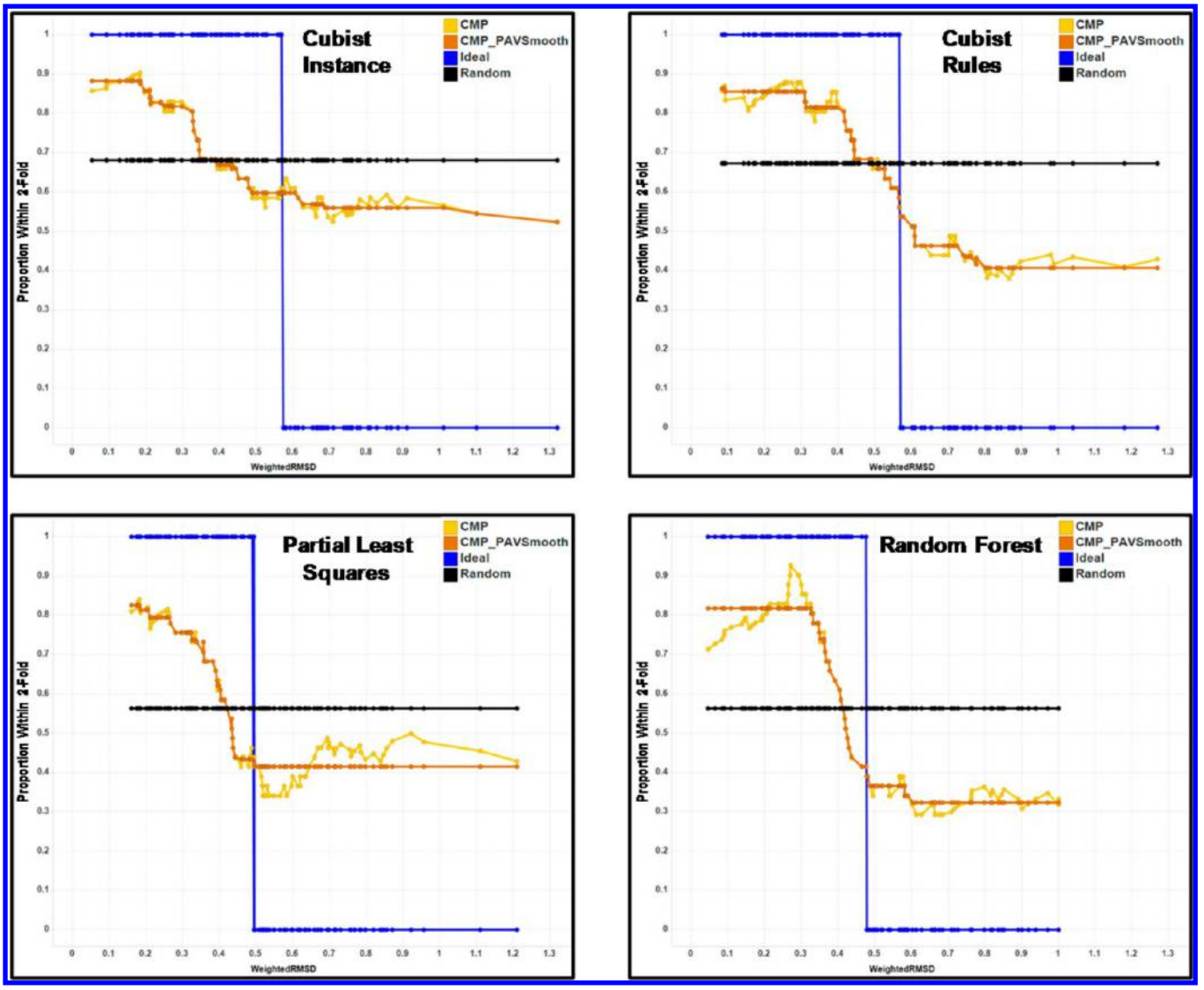

Figure 6. Calibration plots (2-fold) for the Vd_{ss} data set. The black line represents the overall proportion of predictions within 2-fold which is what would be expected if there was no improvement from the confidence metric. The blue line represents the ideal performance for a confidence metric; the case where all of the compounds within 2-fold have the highest confidence. The yellow line is the count moving proportion (CMP; n = 20) of predictions within 2-fold. The orange line is the monotonic pool adjacent violators (PAV) smooth of the yellow line and is the fit used to generate the final confidence metric values.

modeling methods within a data set; however, PLS is the worst performing method in all cases.

Calibration Plots. To understand the general characteristics of the wRMSD values and how they map to the model confidence metric, calibration plots were generated for the four literature test sets (Figures 3−6). In these plots, the *Y*-axis represents the proportion of compounds with good predictions and the *X*-axis is the central wRMSD value for the range of sorted wRMSD values that the proportion was calculated over. The black line represents the probability that a prediction is good for the entire test set. This establishes the baseline or random estimate for the prediction confidence, and it is what one would expect if the confidence metric provided no information. The blue line in the graphs represents the ideal scenario where all compounds with good predictions indeed have the lowest wRMSD values (highest confidence). In this case, the proportion value is 1.0 for compounds with good predictions and 0.0 for compounds with bad predictions. This curve represents the best possible performance of a confidence metric (the upper bound). The yellow line is the RMP or CMP of compounds with good predictions and the orange line is the PAV monotonic smooth of the yellow line.

There are a few general observations one can make from these graphs. First, for the RMP data sets (AIDS, CYP1A2), there are many more compounds with low wRMSD values which translates to very tight agreement between the moving proportion and the monotonic smoothed data. Second, at higher values of wRMSD, where there are fewer data points, the RMP values are highly variable. However, the PAV monotonic smoothing adequately smoothes this variability. Third, for the CMP data sets (DHFR, Vd_{ss}), the wRMSD values are more evenly distributed and there is less variability across the range of values. Finally, and most importantly, in all four literature data sets examined, the confidence metric performs significantly better than baseline (random) for all of the modeling methods.

The threshold for good predictions from the four models built on the literature data was set to be within 2-fold of experiment. For the AIDS models (Figure 3), confidence

dx.doi.org/10.1021/ci300554t | J. Chem. Inf. Model. 2013, 53, 368−383

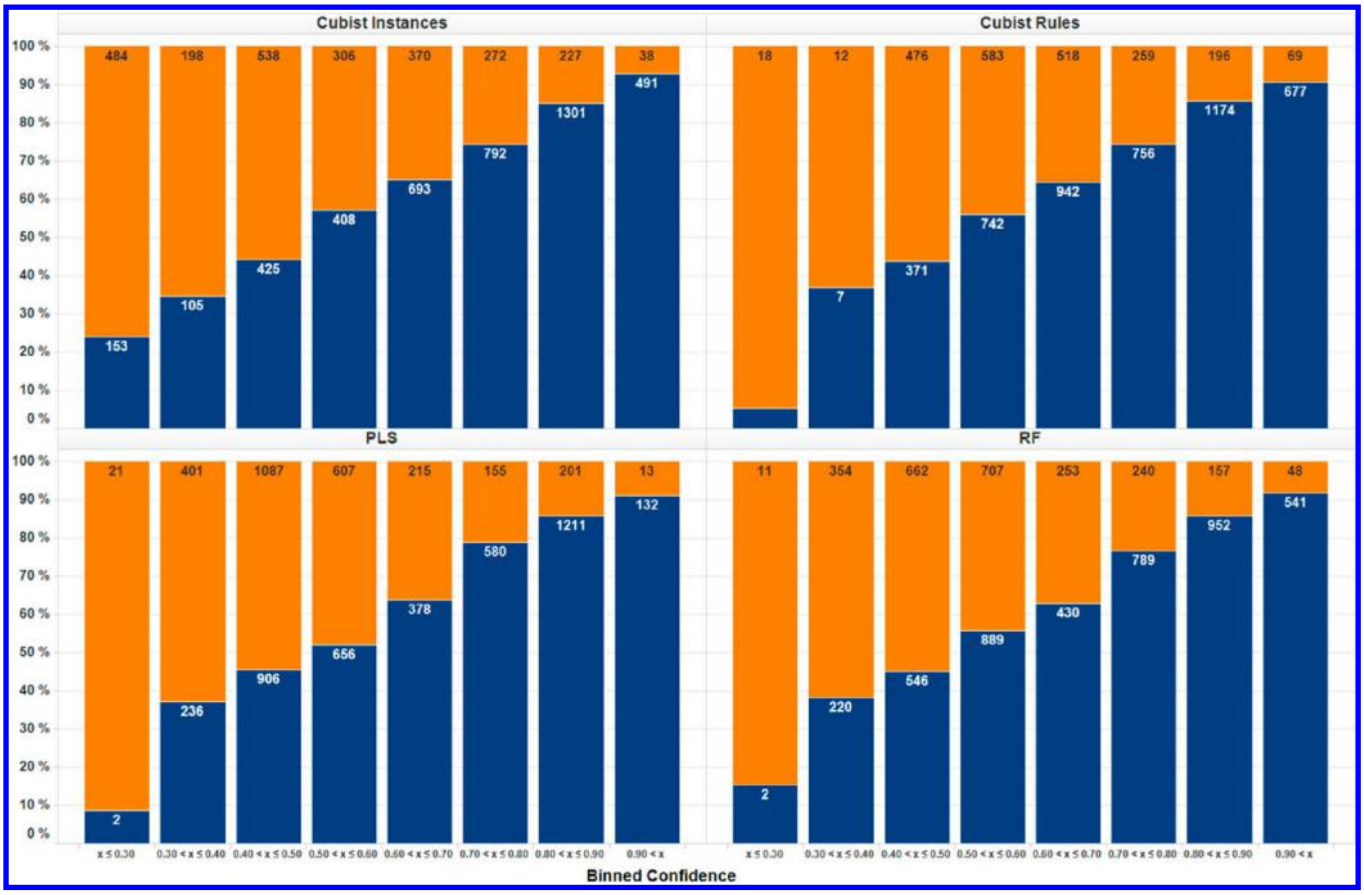

Figure 7. Confidence metric performance for the AIDS validation set for each of the different modeling approaches. Blue segments represent compounds with fold prediction error ≤2, and orange segments are the compounds with fold prediction error >2.

metric values were as high as 0.9 with baseline proportions of 0.60 to 0.69. This was also the case for the CYP1A2 model (Figure 4) where the maximum confidence ranged from 0.87 (PLS) to 0.96 (Cubist rules) with baseline values between 0.38 and 0.46. The confidence metric also worked well on the DHFR models (Figure 5), although the maximum and baseline values were both significantly lower than the other models. The baseline probabilities of being within 2-fold for this model were 0.24−0.40. It is important to point out that even though the wRMSD value range is comparable for the DHFR and AIDS data sets, the calibration step results in significantly different confidence values due to the definition of what is a good or bad prediction. This demonstrates the importance of placing the confidence calculations into the appropriate context for interpretation via calibration. Finally, the Vd_{ss} models (Figure 6) also shows maximum confidence metric values in the 0.8−0.9 range compared to baseline values ranging from 0.56 to 0.68.

Validation Set Performance. To evaluate the performance of the confidence metric, predictions and confidence values were calculated for the validation set, which was completely withheld from the model building process. If the metric performs well, a clear correlation should exist between higher confidence values and the percentage of compounds accurately predicted by the model. In the case of the four models built on literature data, accurate predictions are those that are within 2-fold of experiment. Figures 7−9 and 11 show stacked bar charts of the validation set compounds binned (0.1 increment) by prediction confidence metric values for each of the four modeling methods and end points. The blue segments represent predictions that are within 2-fold of experiment and the orange segments represent prediction errors greater than 2-fold. The numbers within the segments represent the actual number of compounds within that segment.

For the AIDS models (Figure 7), a nice trend exists between the confidence bin ranges and the probability of prediction error within 2-fold. This includes confidence bins above and below the baseline probability range of 0.6−0.7. Also noteworthy is that the percentage of good and bad predictions in each confidence bin is fairly consistent across all four model types. However, the total number of compounds in each bin does vary, indicating differences in model performance. An example of this can be seen by comparing the 0.4−0.5 bin and >0.9 bin across the four methods. What is clear is that more predictions from the PLS model have lower confidence than the other three modeling methods, suggesting that predictions from the PLS model are generally less reliable.

The performance of the confidence metric on the CYP1A2 models (Figure 8) is also good, with some minor differences when compared to the AIDS model. First, there are no confidence values >0.9 for the Cubist instance and PLS models. Second, there are cases when the prediction performance is inconsistent with the confidence values. For example, the Cubist instance model correctly predicts 79% of cases within 2-fold in the 0.6−0.7 confidence bin, which is much better than would be expected. Alternatively, the RF method only predicts 59% of cases within 2-fold in the same confidence bin, which is slightly lower than would be expected. However, there is still

dx.doi.org/10.1021/ci300554t | J. Chem. Inf. Model. 2013, 53, 368−383

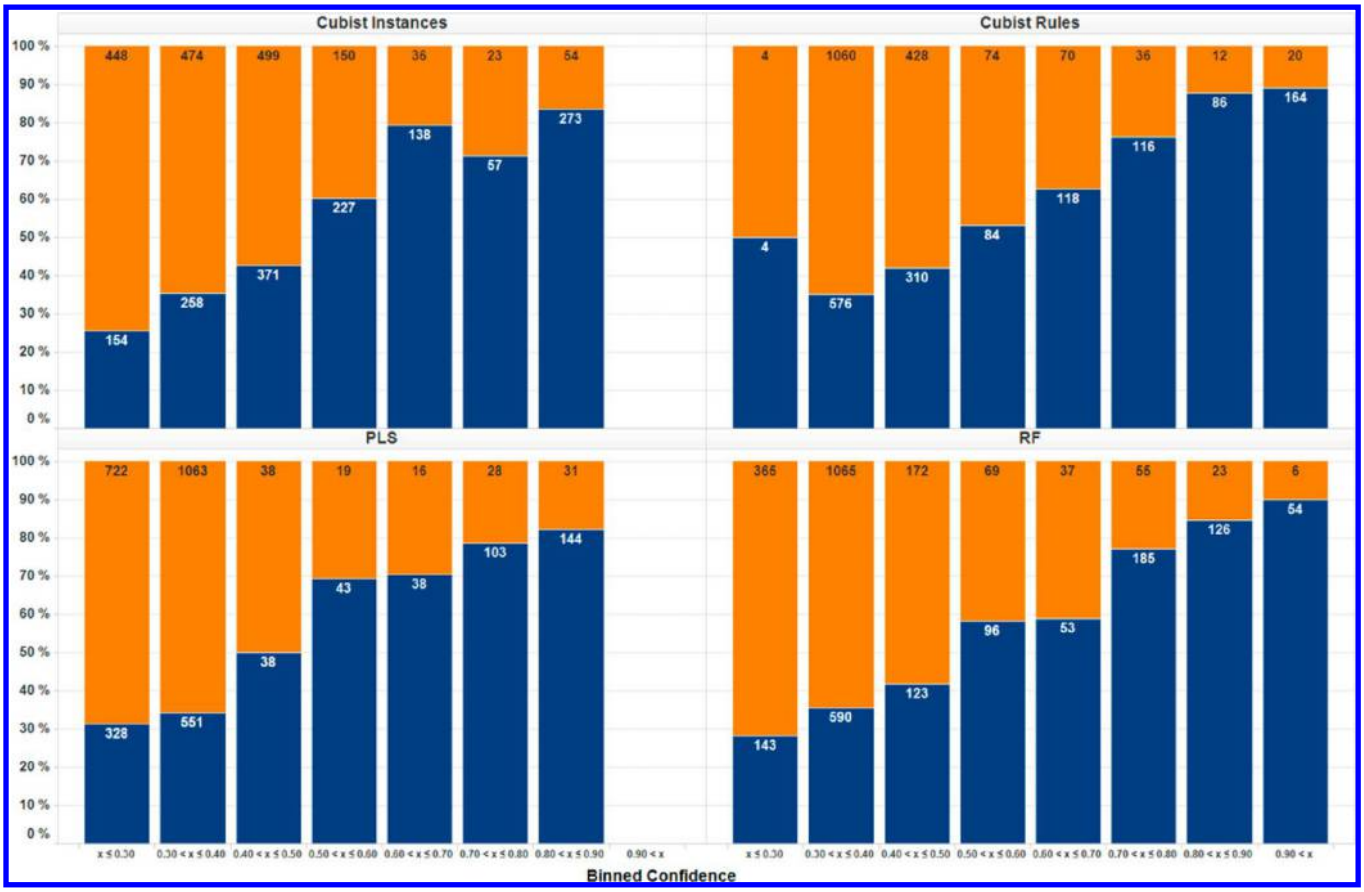

Figure 8. Confidence metric performance for the CYP1A2 validation set for each of the different modeling approaches. Blue segments represent compounds with fold prediction error ≤2, and orange segments are the compounds with fold prediction error >2.

general alignment of the binned confidence values and actual model performance, plus a shifting of compounds to lower confidence bins for the poorer performing PLS method.

Results for the confidence metric performance on the DHFR models (Figure 9) show a very different pattern than for the AIDS and CYP1A2 models. The distribution of confidence values is significantly lower than the other end points with only the RF model having confidence values above 0.50 and there is also less agreement between the predicted confidence and actual performance. A potential cause of the much lower confidence values is the use of 2-fold error for a model that has much higher variability. Figure 10 shows the actual vs predicted DHFR pIC_{50} values for the Cubist instance model colored by the calibrated confidence value. What this illustrates is that although there is a good overall correlation, the number of predictions within 2-fold, denoted by the blue hashed lines, is low. This could be due to model error, but it could also be due to assay variability that is greater than 2-fold. There are significantly more high confidence predictions within 5- or even 10-fold of experimental values. In spite of this, the confidence metric handles this appropriately by assigning lower confidence to these predictions and having fewer, if any, high confidence predictions. This is the benefit of the calibration step—correctly generating an interpretable confidence value that relates to real world performance.

Another potential issue for this assay is the smaller size of the test set (N = 144). Establishing a robust calibration curve is much more difficult when there are fewer samples over which to calculate the moving proportions. Despite this, the confidence values between modeling methods do generally agree and performance differences between them are captured primarily through the number of compounds within the bins (i.e., the ≤0.3 bin in the PLS method contains most of the compounds in its validation set).

Across the four Vd_{ss} models (Figure 11), there is generally good agreement between the proportions of validation set compounds predicted within 2-fold of experiment and the confidence value bins; however, the bin-to-bin variability within a single model is considerable. Much like the DHFR models just discussed, the small size of the Vd_{ss} data set results in very small test and validation sets (N = 119 for each). This, in turn, results in very small sample sizes in the bins inevitably leading to the higher variability. Not unlike the other data sets, the confidence metric does still represent the variation in the performance of different models via the distribution of samples (i.e., the 0.4−0.5 bin contains most of the samples for the PLS method while the 0.8−0.9 bin contains the majority of samples in the better performing Cubist rules model).

Performance of the Confidence Metric for Prospective Data Sets. The best indicator of performance for any new computational tool is its prospective performance on future data. Oftentimes the availability of this data is limited, particularly in academic modeling laboratories, but it can also be difficult to track within industry organizations unless specific processes are in place to capture the data appropriately. As a result, much of the analysis in methodology articles is retrospective in nature and is performed on static, publically available data sets. When constrained by these circumstances,

dx.doi.org/10.1021/ci300554t | J. Chem. Inf. Model. 2013, 53, 368−383

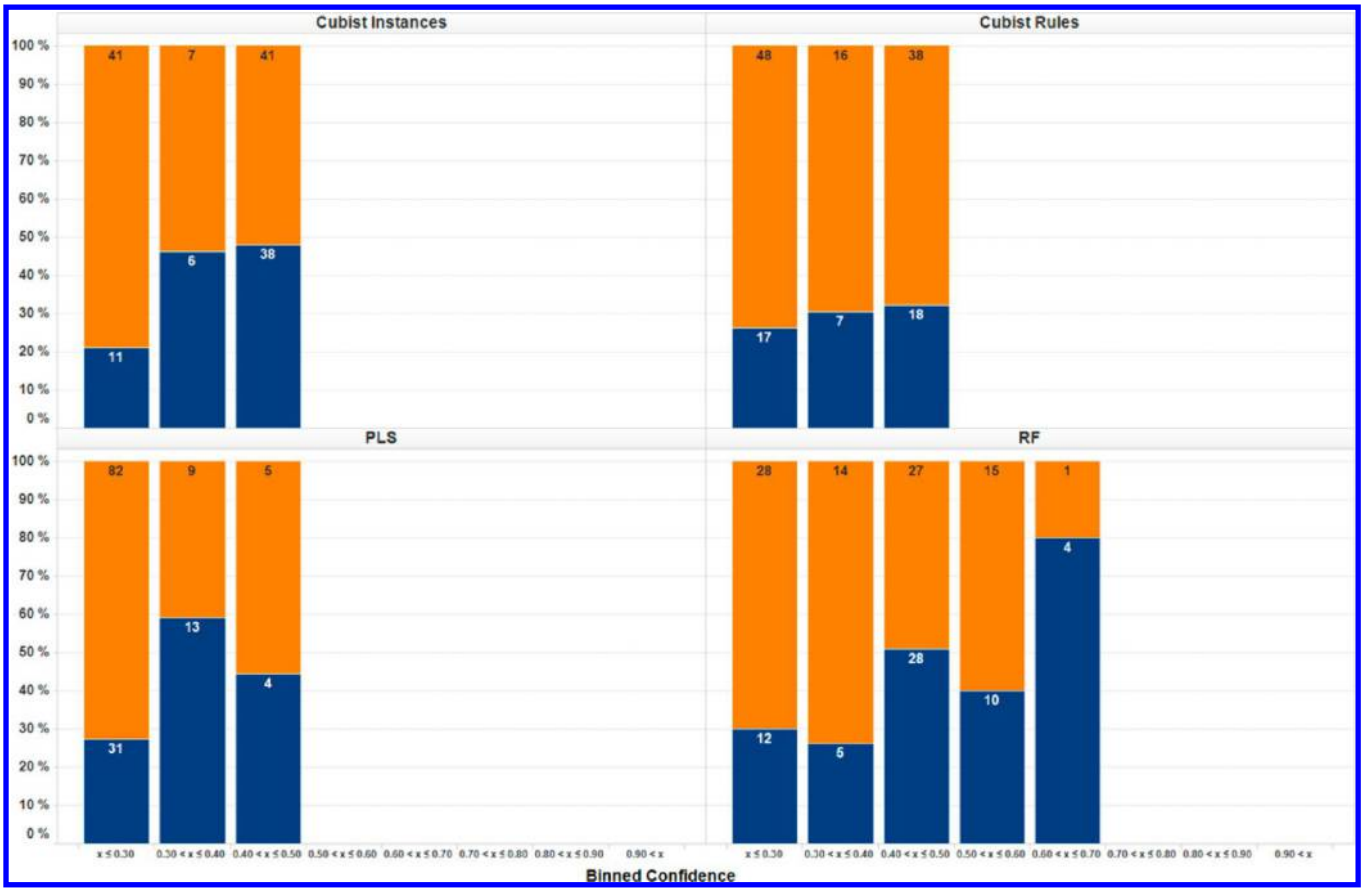

Figure 9. Confidence metric performance for the DHFR validation set for each of the different modeling approaches. Blue segments represent compounds with fold prediction error ≤2, and orange segments are the compounds with fold prediction error >2.

the potential performance of a method is demonstrated by alternative, cross-validation strategies, such as leave-*X*-out, where *X* can be a percentage of compounds, a series or group of compounds, or even just a single compound. When date information is available with a data set, time series simulations can also be a good surrogate for demonstrating prospective performance of a model or method.

To further validate the confidence metric described in this paper, the performance of the confidence metric for five models in our prospective prediction database was evaluated. To generate this database, model predictions and confidence values are captured at the time of compound registration and are not modified even when the models are updated. Once the compounds are tested in our assays, the stored predictions can be compared to the experimental data providing a truly prospective validation. Figure 12a−e show stacked bar charts that capture the percentage of good versus bad predictions, as a function of binned confidence metric values (0.1 increment), for each of the five proprietary models. For all end points except shake-flask logD, the blue segments represent predictions that are within 2-fold of experiment and the orange segments represent prediction errors greater than 2-fold. For shake-flask logD, blue segments represent predictions that are within 0.5 units and orange segments represent prediction errors greater than 0.5 units. In addition, Table 2 includes the Pearson's *R*, RMSE, and percentage of good predictions for each model. In all cases, each of the confidence metric bins is populated with a significant number of compounds allowing for a meaningful interpretation of the trends.

Consistent with the performance of the confidence metric on the two largest literature data sets, AIDS and CYP1A2, a clear trend between good predictions and higher confidence values is observed for the four high-throughput ADME models (Figure 12a−d). The *R* (0.90) of the shake-flask logD model demonstrates that this is generally a very predictive model. In fact, 85% of the 66 115 prospective predictions collected from this model have a confidence value ≥0.7. Also noteworthy is that even at the 0.4−0.5 confidence level bin there is a larger proportion of good predictions than bad ones. This does not necessarily constitute a failure of the method, but may result from scenarios where the predicted logD of a new test case is dissimilar from the experimental logD values of its nearest neighbors in the model training set. Even if the prediction is remarkably accurate, a lower prediction confidence will result from this disconnect. We believe this observation is likely an exception to the rule due to the additive nature of substituent contributions to lipophilicity, and it seems quite reasonable that knowledge outside of the five nearest training set neighbors may frequently influence the prediction of a test case for this particular end point.

For the much larger data sets of HLM, P-gp efflux, and RRCK passive permeability, the prospective performances of the models are not as good as logD, with correlation coefficients of 0.51, 0.50, and 0.55, respectively. One interesting observation that can be seen in Figure 12b−d is the consistency of the proportions of good and bad predictions across confidence bins. However, the overall distribution of predictions at higher and lower confidence varies from model

dx.doi.org/10.1021/ci300554t | *J. Chem. Inf. Model.* 2013, 53, 368−383

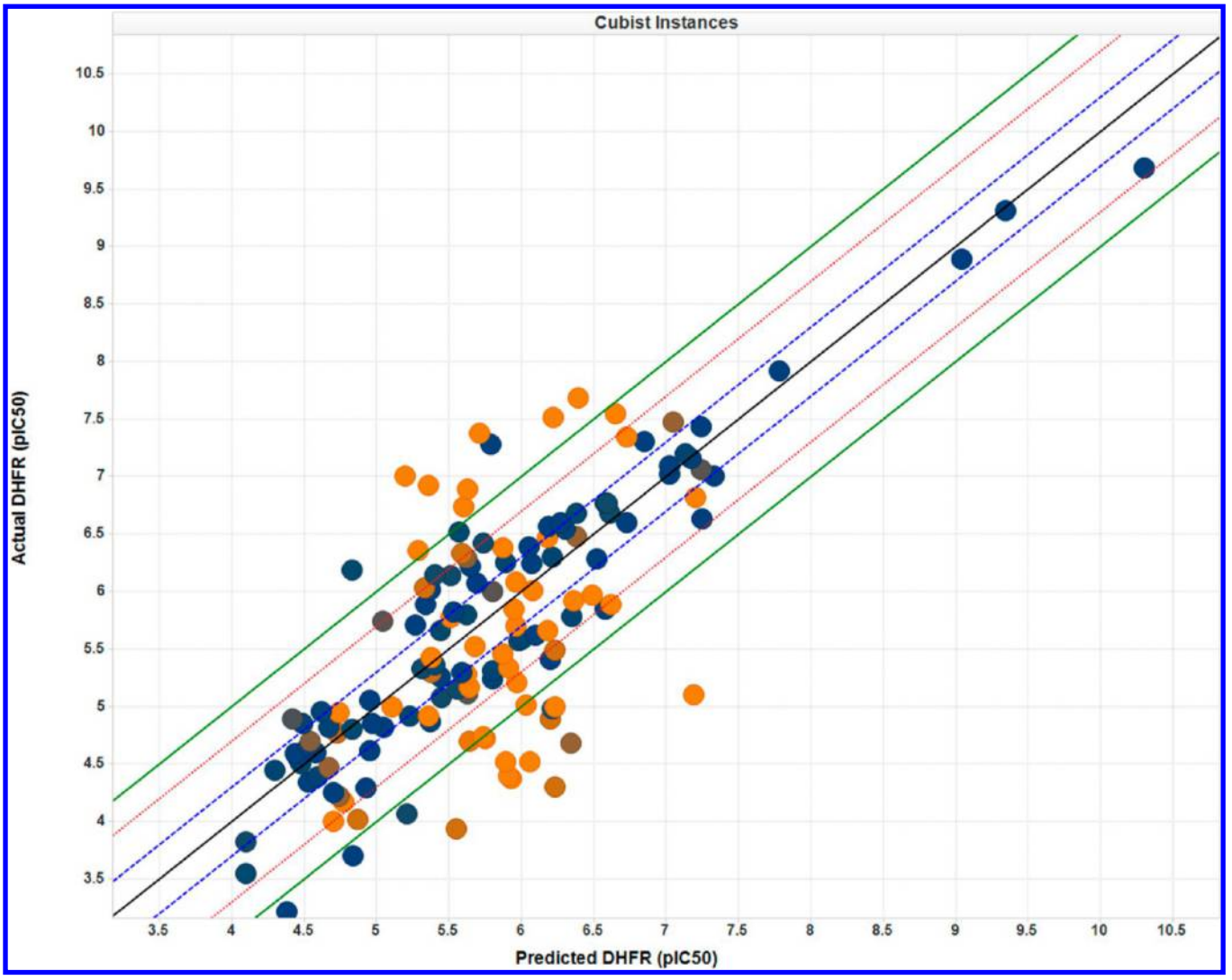

Figure 10. DHFR experimental vs DHFR predicted pIC_{50} values for the Cubist instances model showing large variability. Points are colored by confidence value (blue = high, orange = low). The black line is the line of unity. Blue lines are ±2-fold. Red lines are ±5-fold. Green lines are ±10-fold.

to model. For the HLM model (Figure 12b), only 49% of the predictions have a confidence ≥0.7, while the P-gp (Figure 12c) and permeability (Figure 12d) models have 69% and 58% of predictions above that confidence threshold, respectively. A second observation is that the confidence metric does perform slightly worse for the prospective predictions than for the validation set predictions of the two largest literature data sets, AIDS and CYP1A2. We have limited the comparison to these two literature models since the data set sizes are larger and they were calibrated with the RMP method, consistent with the prospective, in-house models. Comparing the 0.8−0.9 confidence bin in Figures 7 and 8 with the analogous bin in Figure 12a−d, one can see approximately 5−10% reduction in good predictions for the prospective models. The same holds true when comparing the 0.7−0.8, 0.6−0.7, and 0.5−0.6 confidence bins. These trends are expected as prospective data will likely probe new areas of chemical space, whereas even careful selection of validation samples from a static data set will contain some level of chemical homogeneity.

Finally, Figure 12e shows the prospective performance of the confidence metric on a proprietary activity model built upon data for binding (pKI) to a GPCR target. Like the ADME models, a correlation between larger confidence values and a greater percentage of good predictions does exist. Overall, 77% of the prospective predictions are within 2-fold (unlogged values) with a respectable Pearson's *R* correlation coefficient of 0.51 (logged values). Visual inspection of the predicted vs experimental plot (not shown) highlights a general over-prediction of active compounds and an underprediction of inactive compounds, which, in part, explains the suboptimal correlation. Despite this observation, Figure 12e suggests that if one focuses on predictions with confidence values greater than 0.7, a reliable cohort of accurate predictions can be gleaned from the model.

DISCUSSION

Over the past decade, in silico ADMET models have been an integral component in early stage drug design at Pfizer. When classification models were gradually phased out in favor of regression models, the use of similarity-based confidence metrics was cautiously embraced by design teams to understand when predictions were reliable and when they were not. Four years ago, the confidence metric described in this manuscript was developed to improve upon the simplistic, and sometimes

dx.doi.org/10.1021/ci300554t | *J. Chem. Inf. Model.* 2013, 53, 368−383

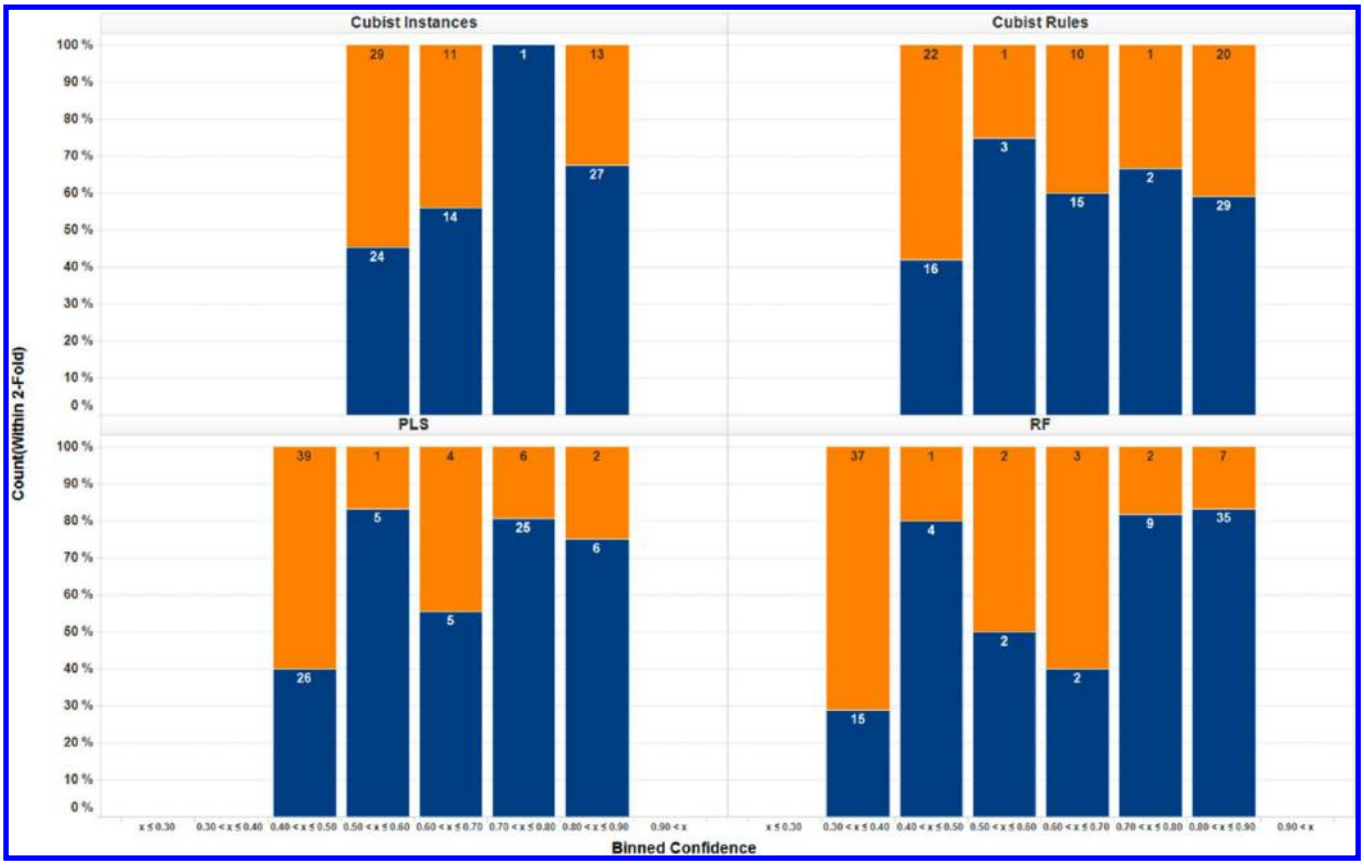

Figure 11. Confidence metric performance for the Vd_{ss} validation set for each of the different modeling approaches. Blue segments represent compounds with fold prediction error ≤2, and orange segments are the compounds with fold prediction error >2.

unreliable, chemical similarity only based methods. Today, this metric is integrated into our in silico ADMET model building process and is used companywide. Through education and experience, design teams have learned how best to use the metric when evaluating the predicted ADMET properties of new compound designs. Over time, some noteworthy observations have been made that the authors feel warrant further discussion.

Choice of Confidence Threshold. The most straightforward use case of any prediction confidence metric would be to simply focus on predictions that have the highest confidence values. In this way, compound designs with favorable ADMET are easy to prioritize over those that are less favorable. When using this confidence metric, one should consider the number of predictions discarded as "low confidence" when setting a confidence level threshold. Table 3 shows the effect of applying incrementally higher confidence metric selection thresholds to the five in-house models. Each row of the table represents a model and each column represents a specific confidence threshold, from 0.7 to 1.0 at 0.05 increments. Each entry of the table contains two pieces of data: the percentage of compounds at or above that confidence threshold and, of those compounds remaining, the percentage of them that are good predictions. The data in Table 3 show that an appropriate choice of confidence threshold to achieve the optimal balance will likely vary from end point to end point. For example, consider the differences between the HLM and P-gp data. At all confidence thresholds evaluated, nearly identical percentages of compounds with good predictions are captured for both models (numbers in parentheses). However, at those same thresholds, 20−30% more P-gp predictions are retained than clearance predictions. To demonstrate this point, consider a project team that is evaluating 100 new ideas in these two models. To ensure that no more than approximately 20% of the predictions are greater than 2-fold, a confidence threshold of 0.85 is used. Based on the historical performance of these models, nearly half of the ideas would meet the confidence threshold from the P-gp model, but only 10% would have sufficiently high confidence from the HLM clearance model. And, one must not forget that approximately 20% of these high confidence predictions will still have greater than 2-fold error based on past performance.

Low Confidence Predictions. Frequently there are a substantial percentage of predictions with low confidence values and the medicinal chemist must decide how to proceed with these ideas. When prediction confidence for a compound is low, it signifies that the activity space of its neighbors is either highly variable or is significantly different from the predicted activity of the new idea. This could mean that the idea is probing a new vector of chemistry space not defined by the model's applicability domain or it could be indicative of a very choppy or clifflike local activity space. Either of these scenarios does not necessarily suggest that the idea should not be followed up on, but that doing so will be done at risk. While in the majority of cases predictions with higher confidence values are most instructive, there are times when a low confidence value may provide useful guidance. For example, a project may elect to use high confidence in silico ADMET predictions in lieu of collecting experimental data, while only submitting compounds that have lower prediction confidence to be screened. This strategy allows teams to still get experimental

dx.doi.org/10.1021/ci300554t | J. Chem. Inf. Model. 2013, 53, 368−383

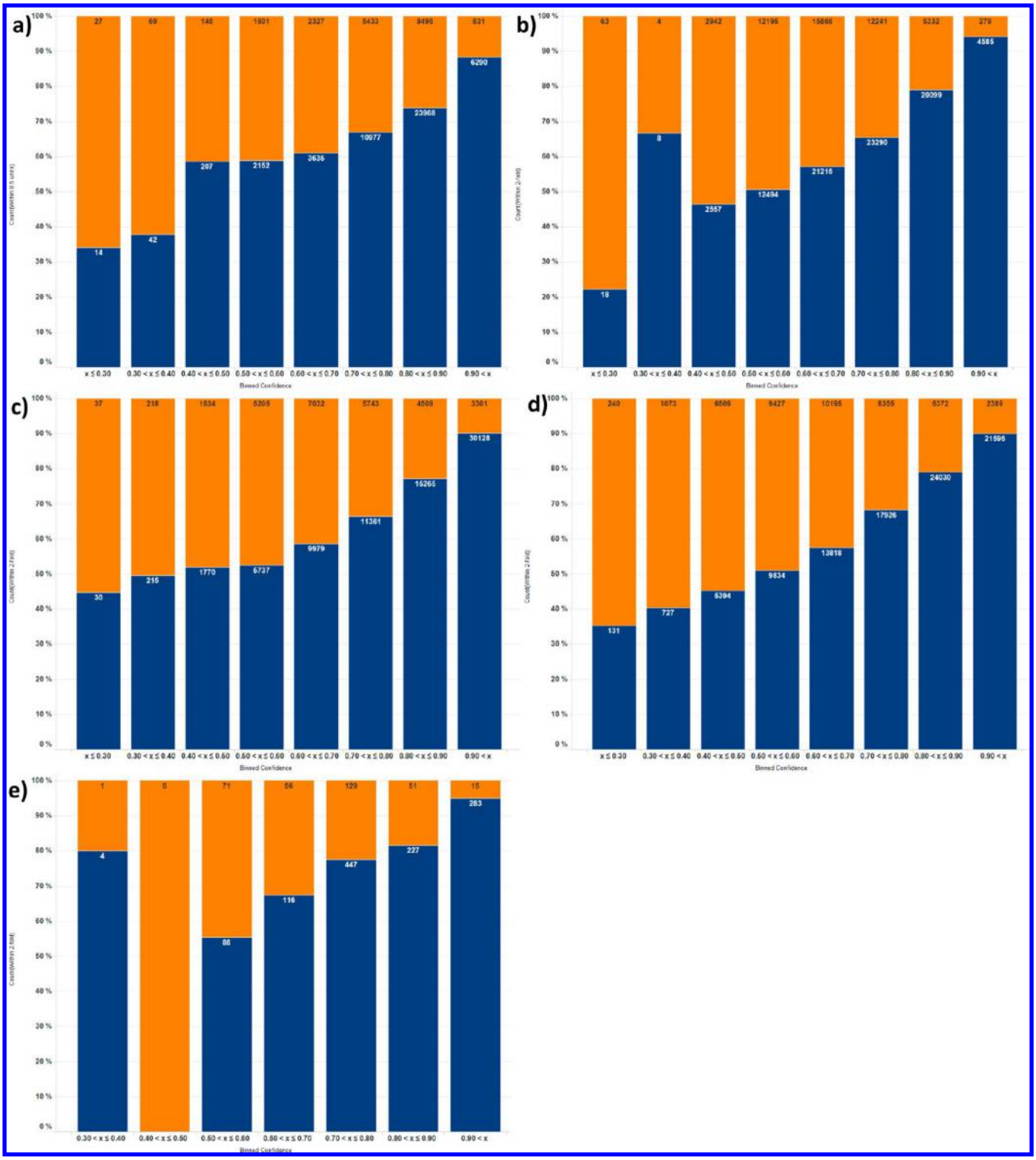

Figure 12. Confidence metric performance for prospective predictions of (a) SFlogD, (b) HLM clearance, (c) P-gp efflux, (d) RRCK permeability, and (e) GPCR binding. Blue segments represent predictions with residual error ≤0.5 (a) or fold error ≤2-fold (b−e) and orange segments represent predictions with residual error >0.5 (a) or fold error >2-fold (b−e).

data on analogues every design cycle but also has the benefit of expanding the applicability domain of the in silico model.

Although it has been demonstrated that this confidence metric works well in its current form, the authors would like to highlight some features where additional scholarship may ultimately improve the method. First, one modification would be to consider alternative, model-appropriate nearest neighbor calculations that do not rely on whole molecule similarity or distance. One proposal would be to use the descriptors chosen by the model method and their associated weightings only, down-weighting or ignoring descriptors that are not relevant to the modeled end point and do not contribute to the predictions. A second proposal might be to use neighbor sets that are inherently defined by the modeling method, such as

dx.doi.org/10.1021/ci300554t | J. Chem. Inf. Model. 2013, 53, 368−383

Table 3. Percentage of Compounds Remaining at Specific Confidence Thresholds and the Percentage of Those Compounds That Have Good Predictions (in Parentheses)

model end point	confidence ≥0.7	confidence ≥0.75	confidence ≥0.8	confidence ≥0.85	confidence ≥0.9
SFlogD	85 (65)	75 (68)	60 (72)	36 (76)	11 (88)
HLM clearance	49 (62)	36 (69)	23 (76)	11 (84)	4 (94)
P-gp efflux	69 (64)	60 (69)	52 (74)	44 (80)	33 (90)
RRCK Permeability	58 (66)	49 (71)	39 (76)	30 (81)	17 (90)
GPCR binding	77 (79)	51 (75)	39 (70)	30 (97)	20 (95)

proximity matrix neighbors for tree based models.[37] A second feature we plan to investigate is activity-range-specific confidence calibration. This stems from the observation that models can have variable accuracy at higher and lower ranges of activity. Currently, the calibration step combines all predictions regardless of predicted activity using the wRMSD values alone. We hypothesize that the confidence performance can be finely tuned by combining predictions that fall within similar activity ranges, especially those at the activity extremes where variability can be high.

CONCLUSION

In this paper, a new approach has been described for calculating and assigning prediction confidence to in silico model predictions. The method considers both the chemical similarity and activity landscape of a compound's nearest neighbors when determining how trustworthy a model prediction is. The foundation of the metric is the value, wRMSD, which combines the predicted value of a compound, the experimental values of its nearest neighbors in the model training set, and the relative distance of those neighbors in the model descriptor space. A straightforward calibration step is used to link wRMSD to the probability of achieving an accurate prediction based on performance for a hold out set, thereby offering an interpretable metric that is useful in triaging model predictions in prospective drug design.

The implementation and performance behavior of the new metric has been clearly demonstrated in the context of four literature data sets of varying size and composition. Through this effort, it has been shown that the approach is broadly applicable across regression-based modeling techniques and can be applied to a variety of end points. Furthermore, we highlighted that the distribution of compounds across binned ranges of confidence are consistent with the overall performance of a modeling method and that there are meaningful differences when calibrating and analyzing the confidence metric for smaller data sets (<1000 compounds) versus larger data sets.

The true utility of the confidence metric was then demonstrated on an expansive set of prospective predictions, captured over a 1−2 year period, from five proprietary Pfizer models. This segment of the analysis further demonstrated that our confidence metric performs robustly in a true drug-discovery setting. The most reassuring result from this prospective analysis was the consistent performance of the metric across all five data sets, both large and small. This supports the use of the metric for models built on both global and project-specific data sets.

Like all methods, this confidence metric has its caveats and nuances. We have highlighted these points, some subtle, others not so much, throughout the manuscript. One example is cases where a high proportion of good predictions are still found at low confidence, such as with the SFlogD model. On the basis of our extensive experience with the method, we can declare this observation to be an exception to the rule. A second example is the use case when low confidence predictions can have practical utility in a project screening strategy. This paradigm can have positive implications for a project, both in bolstering the performance of the model and, after a few design cycles, financially through less experimentation. Finally, a practical point that users must be mindful of is to choose an appropriate confidence threshold above which they can reliably trust, without eliminating too many new ideas. This choice must also be balanced with the performance of the model and confidence metric at different extremes of activity. While these points may not be obvious at the outset of a project, our experience suggests that these behaviors become clearer after only a few design cycles.

In closing, we would like to re-emphasize the value that this confidence metric has added to the drug discovery efforts at Pfizer. As drug discovery becomes increasingly more costly and the pressure mounts to harness and leverage more knowledge computationally, the ability to discriminate good and bad predictions from one another is paramount. What we learned from our research and experience with the nearest neighbor approaches is that there is very likely no one-size-fits-all confidence metric for in silico models. This confidence metric takes a logical step toward bringing together chemical and activity space into a single, interpretable value that impacts decision making in drug design teams.

ASSOCIATED CONTENT

Supporting Information

R Scripts for PLS and random forest model building and prediction as well as lift chart plots for CYP1A2, DHFR, and Vd_{ss} showing the performance of wRMSD compared to chemical distance/similarity confidence in prediction methods. This information is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: christopher.keefer@pfizer.com. Phone: 860-686-4842.

Present Address

⊥Department of Cheminformatics, Early Discovery, Abbott Laboratories, Abbott Park, Illinois 60064, United States.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This research was sponsored by Pfizer, Inc. The authors gratefully acknowledge Matt Troutman and the ADME Technology Group (ATG) for helpful discussions and generation of the in-house ADME data, Greg Bakken in the Computational Sciences Center of Emphasis for technical

discussions on the implementation of the confidence metric, Chris Poss and Steven Heck in the Data and Design Analytics Group for the collection of the prospective model predictions and confidence data, and, finally, Max Kuhn (Nonclinical statistics group), Nate Woody and George Chang (computational ADME group), and Xinjun Hou (World-Wide Medicinal Chemistry) for their thoughtful comments and feedback during the preparation of this manuscript.

REFERENCES

(1) Hansch, C.; Fujita, T. p-σ-π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. *J. Am. Chem. Soc.* **1964**, *86* (8), 1616−1626.
(2) Doweyko, A. Is QSAR relevant to drug discovery? *IDrugs: Invest. Drugs J.* **2008**, *11* (12), 894.
(3) Doweyko, A. M. QSAR: dead or alive? *J. Comput.-Aided Mol. Des.* **2008**, *22* (2), 81−89.
(4) Johnson, S. R. The Trouble with QSAR (or How I Learned To Stop Worrying and Embrace Fallacy). *J. Chem. Inf. Model.* **2007**, *48* (1), 25−26.
(5) Maggiora, G. M. On Outliers and Activity CliffsWhy QSAR Often Disappoints. *J. Chem. Inf. Model.* **2006**, *46* (4), 1535−1535.
(6) Stouch, T. R.; Kenyon, J. R.; Johnson, S. R.; Chen, X. Q.; Doweyko, A.; Li, Y. In silico ADME/Tox: why models fail. *J. Comput.-Aided Mol. Des.* **2003**, *17* (2), 83−92.
(7) Fourches, D.; Muratov, E.; Tropsha, A. Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research. *J. Chem. Inf. Model.* **2010**, *50* (7), 1189.
(8) Dearden, J.; Cronin, M.; Kaiser, K. How not to develop a quantitative structure−activity or structure−property relationship (QSAR/QSPR). *SAR QSAR Environ. Res.* **2009**, *20* (3−4), 241−266.
(9) Scior, T.; Medina-Franco, J.; Do, Q. T.; Martínez-Mayorga, K.; Yunes Rojas, J.; Bernard, P. How to recognize and workaround pitfalls in QSAR studies: a critical review. *Curr. Med. Chem.* **2009**, *16* (32), 4297−4313.
(10) Tropsha, A.; Gramatica, P.; Gombar, V. K. The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR Comb. Sci.* **2003**, *22* (1), 69−77.
(11) Golbraikh, A.; Tropsha, A. Beware of q2! *J. Mol. Graphics Modell.* **2002**, *20* (4), 269.
(12) Sheridan, R. P. Three useful dimensions for domain applicability in QSAR models using random forest. *J. Chem. Inf. Model.* **2012**, *52* (3), 814−23.
(13) Tebby, C.; Mombelli, E. A Kernel-Based Method for Assessing Uncertainty on Individual QSAR Predictions. *Mol. Infor.* **2012**, *31*, 741−751.
(14) Dragos, H.; Gilles, M.; Alexandre, V. Predicting the predictability: a unified approach to the applicability domain problem of QSAR models. *J. Chem. Inf. Model.* **2009**, *49* (7), 1762−1776.
(15) Kühne, R.; Ebert, R. U.; Schüürmann, G. Chemical domain of QSAR models from atom-centered fragments. *J. Chem. Inf. Model.* **2009**, *49* (12), 2660−2669.
(16) Weaver, S.; Gleeson, M. P. The importance of the domain of applicability in QSAR modeling. *J. Mol. Graphics Modell.* **2008**, *26* (8), 1315−1326.
(17) 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* (4), 839−849.
(18) He, L.; Jurs, P. C. Assessing the reliability of a QSAR model's predictions. *J. Mol. Graphics Modell.* **2005**, *23* (6), 503−523.
(19) 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* (6), 1912−28.
(20) Tetko, I. V.; Bruneau, P. Application of ALOGPS to predict 1-octanol/water distribution coefficients, logP, and logD, of AstraZeneca in-house database. *J. Pharm. Sci.* **2004**, *93* (12), 3103−3110.
(21) Quinlan, J. R. In Combining instance-based and model-based learning. *Proceedings of the Tenth International Conference on Machine Learning*, Amherst, MA, June 27−29; Morgan Kaufmann Publishers: San Francisco, 1993; pp 236−243.
(22) Hop, C. E.; Cole, M. J.; Davidson, R. E.; Duignan, D. B.; Federico, J.; Janiszewski, J. S.; Jenkins, K.; Krueger, S.; Lebowitz, R.; Liston, T. E.; Mitchell, W.; Snyder, M.; Steyn, S. J.; Soglia, J. R.; Taylor, C.; Troutman, M. D.; Umland, J.; West, M.; Whalen, K. M.; Zelesky, V.; Zhao, S. X. High throughput ADME screening: practical considerations, impact on the portfolio and enabler of in silico ADME models. *Curr. Drug Metab.* **2008**, *9* (9), 847−53.
(23) National Center for Biotechnology Screening. *PubChem BioAssay Database*; AID=1815, Source=Scripps Research Institute Molecular Screening Center. http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=1815 (accessed October 1, 2011).
(24) http://dtp.nci.nih.gov/docs/aids/aids_data.html (accessed May 1, 2012).
(25) Obach, R. S.; Lombardo, F.; Waters, N. J. Trend analysis of a database of intravenous pharmacokinetic parameters in humans for 670 drug compounds. *Drug Metab. Dispos.* **2008**, *36* (7), 1385−1405.
(26) Sutherland, J. J.; O'Brien, L. A.; Weaver, D. F. Spline-fitting with a genetic algorithm: A method for developing classification structure-activity relationships. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (6), 1906−1915.
(27) Todeschini, R.; Consonni, V.; Mauri, A.; Pavan, M. *DRAGON*, version 6; Talete srl: Milan, Italy, 2011.
(28) Quinlan, J. R. *Data Mining with Cubist.* http://www.rulequest.com/cubist-info.html (accessed October, 2012).
(29) *MoKa*, 1.1.0; Molecular Discovery Ltd.: Perugia, Italy, 2012.
(30) Labute, P. A widely applicable set of descriptors. *J. Mol. Graphics Modell.* **2000**, *18* (4−5), 464−477.
(31) Quinlan, J. R. In Learning with Continuous Classes. *Proceedings of the 5th Australian Joint Conference on Artificial Intelligence*, Hobart, Australia, Nov 16−18; World Scientific: Hobart, Australia, 1992; pp 343−348.
(32) Kuhn, M. Building predictive models in R using the caret package. *J. Stat. Software* **2008**, *28* (5), 1−26.
(33) R Development Core Team. R: A Language and Environment for Statistical Computing. http://www.R-project.org (accessed October, 2012).
(34) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38* (6), 983−996.
(35) de Leeuw, J.; Hornik, K.; Mair, P. Isotone Optimization in R: Pool-Adjacent-Violators Algorithm (PAVA) and Active Set Methods. Department of Statistics Papers; Department of Statistics, UCLA: Los Angeles, 2009; http://www.escholarship.org/uc/item/9zx9c72c (accessed October, 2012).
(36) Barlow, R. E.; Bartholomew, D. J.; Bremner, J.; Brunk, H. *Statistical inference under order restrictions: The theory and application of isotonic regression*; J. Wiley: New York, 1972.
(37) Keefer, C. E.; Woody, N. A. Rejecting unclassifiable samples with decision forests. *Chemom. Intell. Lab. Syst.* **2006**, *84* (1−2), 40−45.