JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

Critically Assessing the Predictive Power of QSAR Models for Human Liver Microsomal Stability

Ruifeng Liu,* Patric Schyman, and Anders Wallqvist*

DoD Biotechnology High Performance Computing Software Applications Institute, Telemedicine and Advanced Technology Research Center, U.S. Army Medical Research and Materiel Command, MCMR-TT, 504 Scott Street, Fort Detrick, Maryland 21702-5012, United States

Supporting Information

ABSTRACT: To lower the possibility of late-stage failures in the drug development process, an up-front assessment of absorption, distribution, metabolism, elimination, and toxicity is commonly implemented through a battery of *in silico* and *in vitro* assays. As *in vitro* data is accumulated, *in silico* quantitative structure−activity relationship (QSAR) models can be trained and used to assess compounds even before they are synthesized. Even though it is generally recognized that QSAR model performance deteriorates over time, rigorous independent studies of model performance deterioration is typically hindered by the lack of publicly available large data sets of structurally diverse compounds. Here, we investigated predictive properties of QSAR models derived from an assembly of publicly available human liver microsomal (HLM) stability data using variable nearest neighbor (*v*-NN) and random forest (RF) methods. In particular, we evaluated the degree of time-dependent model performance deterioration. Our results show that when evaluated by 10-fold cross-validation with all available HLM data randomly distributed among 10 equal-sized validation groups, we achieved high-quality model performance from both machine-learning methods. However, when we developed HLM models based on when the data appeared and tried to predict data published later, we found that neither method produced predictive models and that their applicability was dramatically reduced. On the other hand, when a small percentage of randomly selected compounds from data published later were included in the training set, performance of both machine-learning methods improved significantly. The implication is that 1) QSAR model quality should be analyzed in a time-dependent manner to assess their true predictive power and 2) it is imperative to retrain models with *any* up-to-date experimental data to ensure maximum applicability.

INTRODUCTION

Modern drug discovery is a costly and risky endeavor where failures in late-stage clinical trials still remain common, essentially negating all prior compound development investments.[1,2] Thus, "fail early, fail cheap" is imperative, and multiple *in vitro* assays have been developed to assess pharmacokinetics properties early in the hit/lead generation stage.[3] Among them, the human liver microsomal (HLM) stability assay is perhaps the one most commonly utilized for assessing clearance of chemicals by the human liver—the most important organ for drug metabolism. In a phase I HLM assay, a compound is initially incubated in human liver microsomes containing major drug metabolizing cytochrome P450 (CYP) enzymes. The fraction of compound remaining is then quantified at different time-points by liquid chromatography−mass spectroscopy, and the half-life or clearance rate is calculated from the time-series data.[4] Compounds with short half-lives are quickly metabolized in the liver, and a significantly higher dose may be needed to achieve a desired therapeutic concentration in systemic circulation. However, high dosages are associated with increased drug-induced toxicity, the main factor for a drug to either fail in the development stage or be withdrawn from the market.[5]

To flag compounds that are rapidly metabolized in the human liver, all major pharmaceutical companies have adopted the HLM assay, with the result that tens of thousands of compounds have been evaluated in this system. In addition, computational HLM stability prediction models[6−9] have been developed for virtual screening and guiding compound design. Because complex molecular mechanisms contribute to HLM stability, development of predictive *in silico* models requires a large number of structurally diverse compounds with HLM data as the training set. Unfortunately, the existing large *in vitro* data sets are not routinely shared in the community, restricting the broad development and utilization of *in silico* models for predicting HLM stability. Furthermore, because the data from the training and test/validation sets of the developed models are not routinely disclosed, the true predictive power of the models cannot be independently verified.

However, over the past decade, a significant amount of HLM stability data for drug-like compounds have been disclosed in piecemeal fashion through research publications and collected in publicly accessible databases, such as ChEMBL.[10] With the

Received: May 5, 2015
Published: July 14, 2015

DOI: 10.1021/acs.jcim.5b00255
J. Chem. Inf. Model. 2015, 55, 1566−1575

availability of this data, which currently stands at thousands of compounds, we evaluated different aspects of quantitative structure−activity relationship (QSAR) model performance from this collective data set. In particular, we critically evaluated whether models developed from the currently available data can be used to predict the HLM stability of compounds or compound series that have not yet been synthesized; i.e., what does it take to make prospective *de novo* predictions. We like to point out that prospective model predictability is the subject of recent papers of Sheridan et al.[11,12] and Wood et al.[13] using Merck and AstraZeneca proprietary data sets, respectively. Results of the present study using the publicly available HLM data set support their proposition that time-split method of dividing training and test sets is a more realistic way of simulating prospective predictions.

MATERIALS AND METHODS

The HLM Data Set. We retrieved HLM data from ChEMBL, a manually curated chemical database[10] of bioactive molecules maintained by the European Bioinformatics Institute, on 6 February 2015. To retrieve the data, we used the assay search functionality with the keywords "human liver microsome," which returned 3,392 assays. We filtered compounds tested in these assays by retaining only entries that had half-life ($T_{1/2}$) as the reported bioactivity. This resulted in a list of 4,274 compound entries. We removed 262 entries that either lacked $T_{1/2}$ values or were associated with nonhuman organisms: *Rattus norvegicus* (brown rat) or *Homona magnanima* (moth). In addition, we removed all entries associated with testing glucuronidation (phase II metabolism) with the uridine diphosphate glucuronic acid (UDPGA) cofactor or those that include both reduced nicotinamide adenine dinucleotide phosphate (NADPH) and UDPGA as cofactors. We further removed data associated with human liver supersomes (microsomes prepared from insect cells infected by baculovirus and containing the complementary DNA of a single human CYP isoenzyme). We also removed data derived from experiments with the presence of inhibitors of some specific CYP isoforms, and data taken in the absence of the NADPH cofactor. The resulting data set contained compound stability information, as measured by $T_{1/2}$, against phase I metabolism in human liver microsomes. However, not all entries in the data set are from structurally unique compounds, as some of the compounds were tested and/or reported multiple times and appeared in data sets from different sources.

We next processed the $T_{1/2}$ data to classify compounds as stable or unstable by selecting a threshold value. In a couple of publications on the same subject, an intrinsic clearance rate (CL_{int}) of 20 mL/min/kg was chosen to classify compounds as stable (CL_{int} < 20 mL/min/kg) or unstable (CL_{int} ≥ 20 mL/min/kg) based on the CL_{int} distribution of a large number of Pfizer compounds (~15,000).[7,8] Using the equation proposed by Obach et al.[14] relating CL_{int} and *in vitro* $T_{1/2}$ as

$$T_{1/2} = \frac{0.693}{CL_{int}} \times \frac{\text{mL incubation}}{\text{mg microsomes}} \times \frac{\text{45 mg microsomes}}{\text{g liver}} \times \frac{\text{21 g liver}}{\text{kg b.w.}} \quad (1)$$

we calculated the corresponding $T_{1/2}$ to be 32.7 min. We thus chose a $T_{1/2}$ of 30 min to categorize a compound as stable ($T_{1/2}$ > 30 min) or unstable ($T_{1/2} \leq$ 30 min) in an HLM assay.

Based on the $T_{1/2}$ threshold of 30 min, we removed from the data set nine compounds that we could not unambiguously classify as stable or unstable; e.g., a reported value of $T_{1/2}$ > 15 min cannot be classified in our scheme. Furthermore, 145 compounds appeared 319 times in the data set. For a compound with multiple entries, we retained a single entry if all their entries had an unambiguous $T_{1/2}$ value either higher or lower than 30 min. Otherwise, we removed all the entries associated with the compound.

The final assembled HLM data set comprised 3,654 compounds, 2,313 (63%) of which were classified as stable, and 1,341 (37%) of which were classified as unstable. Most of the collected data appeared in peer-reviewed journals published from 1998 to 2014; however, 37 entries were not associated with any date. The data set is provided as Supporting Information. We like to point out that all biological assay results are associated with certain degrees of uncertainty. Because the HLM data collected in ChEMBL were derived from different laboratories over a long time, data reproducibility may be lower than expected. We tried to estimate uncertainty of the half-life data from the 319 entries associated with 145 compounds but without success. This was because many of the entries reported $T_{1/2}$ larger or smaller than some threshold values. The estimation was also confounded by that some compounds were reported with exactly the same $T_{1/2}$ in different papers, not representing true multiple determinations, but rather, rereporting of a single measurement. Even though we were not able to give a reliable estimate of the uncertainty of the $T_{1/2}$ in the data set, we believe the data set is noisier than HLM data derived from a single laboratory. Because of higher than expected noise in the data set, we attempted to develop and evaluate HLM stability classification models instead of numerical half-life prediction models. Classification is less susceptible to data variability, because a compound with a numerically different half-life of 60 or 80 min classifies it into the same category (if the classification threshold is 30 min). However, for compounds with half-life close to the threshold, assay reproducibility is still an issue, as different measurements can easily designate a compound in different categories. Considering that all biological assays have reproducibility issues and therefore a perfect training set does not exist, we feel that the goal of QSAR modeling should not be producing a model that gives perfect prediction of the training set, because doing so will inevitably lead to overfitting.

Molecular Descriptors. Because multiple CYP enzymes as well as some other enzymes contribute to the metabolism of chemicals in a phase I HLM assay, the molecular mechanisms by which the compounds are metabolized are complex and largely unknown. For such systems, the only basis for computational predictions is the principle of similar molecular structures having similar bioactivities. Thus, the appropriate molecular descriptors should provide detailed descriptions of molecular structures. Molecular fingerprints provide the most detailed descriptions of molecular structures, and they are widely used in chemical database searches for structurally similar compounds. In this study, we used the extended connectivity fingerprint (ECFP)[15] from Biovia as the molecular descriptor. The fingerprint was generated iteratively to encode features that represent each atom in larger and larger structural neighborhoods. At iteration 0 (ECFP_0), we encoded the information on individual atoms by turning on a corresponding bit in a binary bit string. The information includes the number of connections (bonds) to the atom, element type, charge, and

DOI: 10.1021/acs.jcim.5b00255
J. Chem. Inf. Model. 2015, 55, 1566−1575

atom mass. At iteration 1, we encoded the information on all atoms directly bonded to the atom (within a diameter of two chemical bonds and, hence, termed ECFP_2). At iteration 2 (ECFP_4), we encoded the information on all atoms within a diameter of four chemical bonds. When we reached the desired neighborhood size, the process was complete, and the set of bits representing all features of the atom was returned as part of the molecular fingerprint. We repeated this process for all the atoms in a molecule. The molecular ECFP_n fingerprint is, thus, a collection of all the bits representing atoms in their molecular neighborhoods, where each bit represents a specific molecular structure moiety and is called a bit-feature.

With increasing n, ECFP_n gives an increasingly more detailed description of the molecular structure. However, with increasing n, the number of unique bit-features increases exponentially, and so does the computational cost. A practical strategy to balance the cost and performance is to fold an original fingerprint into a fixed-length bit string by the logical OR operation.[16] The folding leads to loss of information due to bit-feature clashing, with the degree of information loss proportional to the degree of folding. To balance the computational cost and accuracy of molecular structure description, we chose to use an ECFP_4 fingerprint folded to a fixed length of 2,048 bits as molecular descriptor in this study. The pros and cons of using ECFP_4 versus a larger fingerprint size such as ECFP_6 were discussed in previous studies.[11] Generally speaking, ECFP_4 gives a reasonably good description of molecular structure details. Larger fingerprint size provides even more detailed descriptions but at the expense of significantly longer bit strings. If the bit strings have to be folded to a computationally manageable length, information loss due to folding may cancel the benefits of a larger fingerprint size.

Random Forest Method. Random forest (RF)[17,18] is one of the most popular machine-learning methods used in the chemoinformatics/QSAR community. In previous publications of *in silico* prediction of HLM stability, RF was found to outperform several other machine-learning methods, including support vector machines, logistic regression, recursive partitioning, and a naïve Bayesian classier.[7,8] We therefore selected RF as our first method in this study. To develop an RF model, we trained 500 decision trees. Each of them used a subset of ECFP_4 bit-features to recursively partition the training set samples so that the stable and unstable compounds were enriched in different branches. To predict the HLM stability of a test compound, we used all 500 decision trees. A compound was categorized as unstable if it was predicted unstable by more than 50% of the trees. We used the RF module of the R Project for Statistical Computing[19] as implemented in Pipeline Pilot.[20]

Variable Nearest Neighbor Method. Similarly, based on the premise of similar structures having similar activities, the k-nearest neighbor method should be well suited for QSAR, as it always uses k-nearest neighbors to make a prediction. As such, it is used in a number of studies of metabolic stability.[9,21] A shortcoming of this method is that it always bases a prediction on a constant number of nearest neighbors, irrespective of whether the nearest neighbors are structurally similar enough to ensure similar activity. To correct for this shortcoming, we have proposed a variable nearest-neighbor (v-NN) method.[16] Instead of using a constant number of nearest neighbors, v-NN uses all nearest neighbors meeting a structural similarity criterion for making a prediction. When no nearest neighbor meets the similarity criterion, we do not make a prediction in order to maintain the overall prediction reliability. In essence, the predicted property y is a weighted average across structurally similar neighbors, as

$$y = \frac{\sum_{i=1}^{v} y_i e^{-\left(\frac{d_i}{h}\right)^2}}{\sum_{i=1}^{v} e^{-\left(\frac{d_i}{h}\right)^2}}, \quad d_i \le d_0 \tag{2}$$

where d_i denotes the Tanimoto distance between a target molecule for which a prediction is made and molecule i of the training set. The Tanimoto distance was defined as 1 − TC, where TC is the Tanimoto similarity coefficient between two the molecules. y_i denotes the experimentally measured value of molecule i; v denotes the total number of training set molecules satisfying the condition $d_i \le d_0$. h is a smoothing factor which dampens the distance penalty, and d_0 is a Tanimoto-distance threshold beyond which two molecules are not considered sufficiently similar to include in the average. To predict HLM stability, we assigned a y_i value of 1 to all unstable compounds and a value of 0 to all stable compounds in the HLM data set. Using eq 2, the predicted HLM stability value falls between 0 and 1. A value below 0.5 classifies a compound as stable; otherwise, a compound is classified as unstable.

We used both RF and v-NN methods in the present study based on the following considerations. First, RF was one of the top-performing methods for HLM stability prediction in previous studies, and therefore it should be included in the present study as a benchmark. Second, v-NN is a recently developed method that needs to be critically evaluated and compared to existing standard models. Third, RF and v-NN use different aspects of molecular structure information. RF uses presence/absence of specific molecular structural fragments to partition the samples into different subsets and thus focuses on individual molecular structural features. v-NN, on the other hand, uses overall molecular structural similarity as measured by Tanimoto distance. Thus, the two methods may complement each other, and a consensus prediction may be more reliable than either method alone. We used both methods in the study to evaluate if they indeed complement each other.

The RF and v-NN methods have two major advantages compared to regression-based modeling approaches. First, they can handle the situation of multiple molecular mechanisms contributing to an experimental end point. In the RF approach, each decision tree or a subset of decision trees may represent contribution of a distinct molecular mechanism. In the v-NN, a prediction is made from information on structurally similar compounds only. The presence of structurally dissimilar compounds in the training set that may be associated with distinct molecular mechanisms do not affect v-NN predictions, as structurally dissimilar compounds are excluded from distance-weight-average. To handle the same situation with multiple linear regression or similar approaches, multiple equations may need to be established, each of them may represent a distinct molecular mechanism. Second, both v-NN and RF are less susceptible to overfitting. Even though we used 2,048 ECFP_4 fingerprint bit features as molecular descriptors, there are only two adjustable parameters in the v-NN method, the Tanimoto distance threshold and the smoothing factor, that need to be determined from the training set data. This is in sharp contrast to many other methods that require at least as many experimental data points in the training set as the number of molecular descriptors. With the RF approach, we used a subset (∼45) of the 2,048 bit-features to build each decision

DOI: 10.1021/acs.jcim.5b00255
J. Chem. Inf. Model. 2015, 55, 1566−1575

tree. At each branching point, a statistical test was performed. A bit-feature that gave no enrichment of positive and negative samples in different leaf nodes was skipped, and another bit-feature was tested. Bit-features that are present in very small number of training set samples did not survive the statistical tests, as they could not provide statistically significant enrichment of the positive and negative samples due to small sample size. Thus, overfitting was minimized.

Model Performance Measures. We used the following metrics to measure the quality of the classification models

$$sensitivity = \frac{TP}{TP + FN} \quad (3)$$

$$specificity = \frac{TN}{FP + TN} \quad (4)$$

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (5)$$

$$kappa = \frac{accuracy - Pr(e)}{1 - Pr(e)} \quad (6)$$

where TP is true positive, TN is true negative, FP is false positive, and FN is false negative. Kappa is a metric for assessing the quality of binary classifiers,[22] and $Pr(e)$ is an estimate of the probability of a correct prediction by chance. It is calculated as

$$Pr(e) = \frac{(TP + FN)(TP + FP) + (FP + TN)(TN + FN)}{(TP + TN + FP + TN)^2} \quad (7)$$

In essence, kappa compares the probability of correct predictions to the probability of correct predictions by chance. Its values range from +1 (perfect agreement between model prediction and experiment) to −1 (complete disagreement), with 0 indicating no agreement above that expected by chance. As a good measure of the quality of a binary classifier, kappa's merit over accuracy is easy to appreciate with an imbalanced data set, e.g., a data set in which 90% of the samples belong to one class and the remaining 10% of samples belong to another class. A meaningless classifier that simply assigns everything to the majority class would have a decent accuracy of 90% for such a data set, as no more than 10% of the samples would be incorrectly assigned. For such a data set, kappa of the meaningless classifier would be 0, as $Pr(e)$ of the meaningless classifier would be 90%.

In addition to the above metrics, we also considered coverage—the percentage of samples within the applicability domain for a given data set—as a performance measure. After all, a model offers little practical value if it has a very small applicability domain, even if it can give perfect predictions for a very small number of samples.

RESULTS AND DISCUSSION

Performance of *v*-NN as Evaluated by 10-Fold Cross-Validation. With the *v*-NN method, there are two adjustable parameters that may influence performance: the molecular structural similarity threshold d_0 and the smoothing factor h in eq 2. To determine the optimal values of these parameters, we divided the HLM data set randomly into 10 equal-sized groups for 10-fold cross-validation. We used nine groups as a training set for model development and predicted the HLM stability of the excluded group, repeating this process until each and every group was left out once for the evaluation of model performance. The Supporting Information provides the membership information on each compound in the cross-validation groups. We determined the optimal Tanimoto distance threshold d_0 and smoothing factor h via a series of 10-fold cross-validation calculations by increasing h stepwise from 0.1 to 1.0 (step size 0.1) and increasing d_0 stepwise from 0.05 to 0.75 (step size 0.05). For the data set, we found that a d_0 of 0.45 and an h of 0.20 achieved a good balance of prediction reliability and model coverage. Figure 1(a) shows

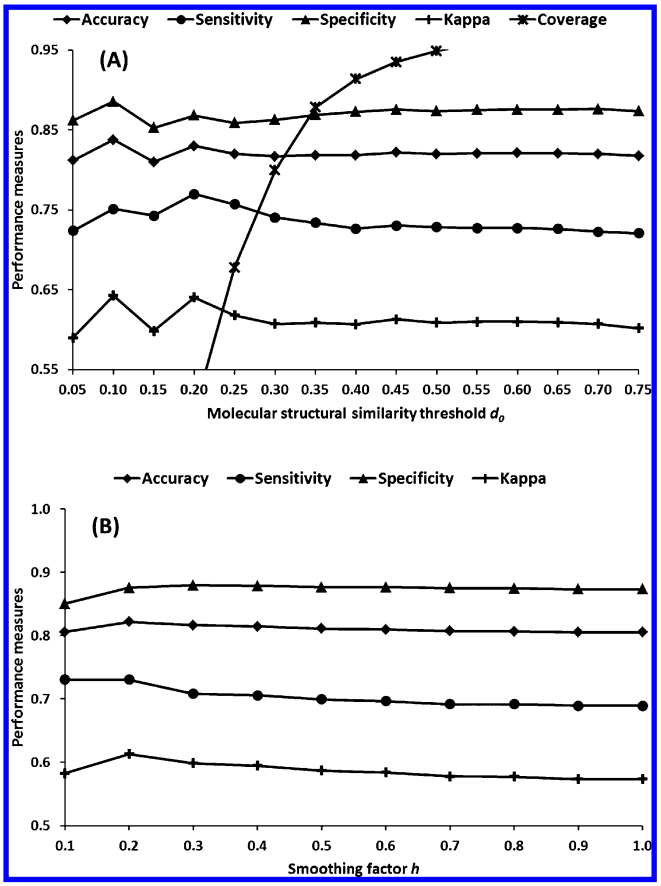

Figure 1. (a) Performance measures of variable nearest neighbor (*v*-NN) method with respect to Tanimoto distance threshold d_0 at a constant smoothing factor h of 0.20. (b) Performance measures of *v*-NN with respect to smoothing factor h at a constant Tanimoto distance threshold d_0 of 0.45. The coverage (94%) is not shown, as it is constant with a d_0 of 0.45.

the model performance measures versus d_0 obtained at a constant smoothing factor of h = 0.20. With low d_0 values, model performance as measured by accuracy and kappa could be high, but the coverage was very low, meaning that a majority of the compounds do not have neighbors meeting the stringent molecular structural similarity requirement. Note that at extremely low d_0 values, model performance was highly variable. This was because the model coverage was extremely low (a very small number of molecules had prediction results) and, therefore, the performance measures were statistically unreliable. With increasing d_0, model performance deteriorated gradually, whereas model coverage increased significantly. With a d_0 of 0.45, the model had a coverage of 94%, accuracy of 82%, sensitivity of 73%, specificity of 88%, and a kappa value of 0.61.

Figure 1(b) shows the influence of the smoothing factor h on model performance obtained at a fixed d_0 of 0.45. The coverage

DOI: 10.1021/acs.jcim.5b00255
J. Chem. Inf. Model. 2015, 55, 1566−1575

is not shown, because at a constant d_0 of 0.45, the coverage is constant at 94%. The figure shows that the specificity remained above 85% across different h values, and sensitivity stayed around 70%. The disparity between sensitivity and specificity is common for imbalanced data sets (data sets with one class having significantly more samples than the other does). The accuracy and kappa values indicated that an h of 0.20 provided the best performance for v-NN with d_0 = 0.45.

In summary, v-NN with d_0 = 0.45 and h = 0.20 provided good performance when evaluated by 10-fold cross-validation with all compounds randomly distributed into 10 equal-sized groups. It achieved an overall accuracy of 82% with a kappa value of 0.61 for 94% of the compounds. For the other 6% (237) compounds, the model failed to make a prediction, because no training set compounds were within the trusted Tanimoto distance of 0.45. For comparison, we made v-NN predictions for these compounds by extending d_0 to 0.75 while maintaining h at 0.20. This allowed predictions for 216 of the 237 compounds with an overall accuracy of 74%, a sensitivity of 52%, a specificity of 84%, and a significantly lower kappa value of 0.37. A sensitivity of 52% indicates that nearly half of the unstable compounds were predicted as stable, which was a major contributor to the low kappa value. The kappa plot in Figure 1(a) does not reflect model performance deterioration adequately, as it shows an almost plateaued behavior at high d_0. This was because more than 94% of the compounds have qualified near neighbors, and their HLM stability was predicted satisfactorily. The small number of compounds that were predicted poorly had a negligible negative impact on the overall performance when predictions for all the compounds were considered together.

Performance of RF as Evaluated by 10-Fold Cross-Validation. We first performed 10-fold cross-validation of the RF method without considering the applicability domain, i.e., without consideration of whether the method could give a reliable prediction. The performance of the RF model for the HLM data set was excellent. The overall accuracy, sensitivity, specificity, and kappa were 82%, 77%, 85%, and 0.61, respectively. Because the RF recursively partitions samples based on the presence/absence of specific structural moieties (ECFP fingerprint features), any structural moieties of a test compound that were absent from the training set compounds may pose a challenge to the reliability of the RF predictions. That is, the number of ECFP fingerprint features of a test compound absent from the training set compounds may inversely correlate with RF prediction accuracy. This was shown to be the case in our previous RF study of AMES mutagenicity.[23] To test whether it also applied to the HLM data set, we repeated the 10-fold cross-validation while tracking the number of ECFP_2 fingerprint features missing in the training set. We then calculated model performance measures separately for compounds without any structural moieties absent from the training set and for compounds with an increasing number of structural moieties absent from the training set. The results presented in Table 1 and schematically in Figure 2 show clearly that the RF model performance deteriorates with an increasing number of ECFP_2 bit features absent from the training set. The best performance was achieved for compounds with all their ECFP_2 bit features present in the training set molecules. Consequently, we defined the applicability domain of the RF method for the HLM data set as zero ECFP_2 fingerprint features absent from the training set compounds. Results of the 10-fold cross-validation

Table 1. Performance Measures of the RF Model with Respect to the Number of ECFP_2 Bit-Features Not Present in the Training Set Compounds

ECFP_2 features[a]	accuracy	sensitivity	specificity	kappa	fraction[b]
0	0.82	0.78	0.85	0.62	0.921
1	0.77	0.68	0.83	0.51	0.061
2	0.69	0.58	0.83	0.40	0.013
3	0.73	0.25	1.00	0.30	0.003
4	0.00	ND[c]	0.00	0.00	0.000
5	0.67	0.00	1.00	0.00	0.001
6	1.00	ND[c]	1.00	ND[c]	0.000

[a]Count of ECFP_2 fingerprint features in test compounds not present in the training set. [b]Fraction of test set compounds with specific number of ECFP_2 bit features missing in the training set. [c]Not defined due to division by zero.

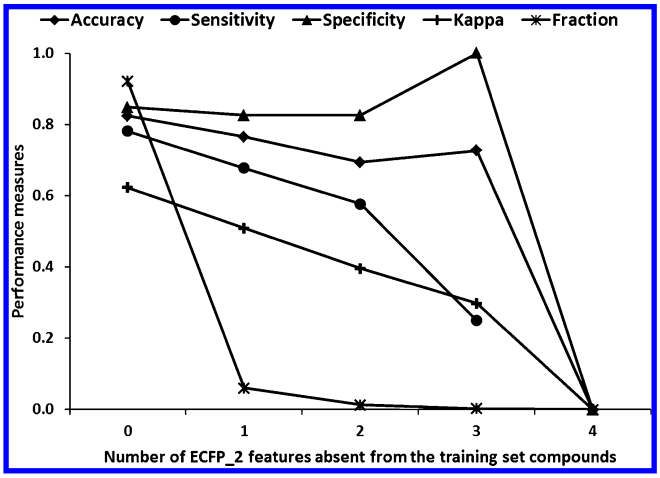

Figure 2. Performance of random forest with respect to the number of missing ECFP_2 fingerprint features in the training set. *Fraction* denotes the fraction of compounds in a category.

calculations indicated that the RF model had a relatively high coverage of 92%.

Based on performance measures derived from the 10-fold cross-validation calculations, the RF outperformed the v-NN slightly when all the samples were randomly distributed into the 10 equal-sized validation groups. This was in agreement with published studies that found that RF generally outperforms most other machine-learning methods.[7,8]

Since ECFP_4 bit-features were used as descriptors in the RF approach, the number of missing ECFP_4 bit-features absent from the training set also correlated with model performance and could thus be used to define the applicability domain. In comparison to ECFP_2 bit-features, ECFP_4 bit-features contain all ECFP_2 bit-features (by definition) as well as considerably larger bit-features spanning up to four chemical bonds. Thus, an applicability domain defined by zero missing ECFP_2 bit-features would potentially correspond to having zero to a few missing ECFP_4 bit-features, i.e., the latter description would be more specific. We prefer defining the applicability domain by the number of missing ECFP_2 bit-features for two reasons. First, ECFP_2 bit-features represent smaller and relatively "simpler" structural fragments. This is important when extending the applicability domain with experimental data, as testing fewer compounds would be required to generate data to retrain the model compared to an ECFP_4-defined applicability domain. Second, metabolic

DOI: 10.1021/acs.jcim.5b00255
J. Chem. Inf. Model. 2015, 55, 1566−1575

stability of many compounds are determined by the presence of soft spots, usually weak chemical bonds that are adequately described by ECFP_2. This is especially true of compound stability with respect to metabolic reactions catalyzed by CYP3A4—the most important drug-metabolizing enzyme in human liver.[24] Thus, the smaller ECFP_2 bit-features represent an adequate and robust choice for defining the applicability domain compared to the more specific ECFP_4 bit-features.

Performance of *v*-NN and RF Consensus Predictions. The *v*-NN and the RF methods base their predictions on different structural aspects of the molecules: *v*-NN uses overall molecular structural similarity, while RF focuses on the presence/absence of individual structural features. We hypothesized that when predictions given by the two methods agree, the chance for these predictions to be correct is higher. To test this hypothesis, we examined the results of 10-fold cross-validation with both methods. We found that 3,206 out of 3,654 compounds (88%) were in the applicability domains of both methods, and predictions by the two methods were in agreement for 2,974 (81%) of the compounds. Data in Table 2 show that consensus predictions by the two methods were indeed more reliable, as the overall accuracy reached 85% and the maximum kappa value was 0.68, significantly higher than the kappa values of either method alone. For the 232 compounds for which the two methods gave different predictions, neither method performed well, as both of them had an overall accuracy close to 50% and kappa values close to 0 (Table 2).

Table 2. Performance of 10-Fold Cross-Validation of RF, *v*-NN, and RF/*v*-NN Consensus Predictions, as Well as RF and *v*-NN Nonconsensus Predictions

model	accuracy	sensitivity	specificity	kappa	coverage
RF	0.82	0.78	0.85	0.62	0.92
v-NN	0.82	0.73	0.88	0.61	0.94
Consensus[a]	0.85	0.79	0.89	0.68	0.81
RF[b]	0.52	0.86	0.25	0.10	N/A[c]
v-NN[d]	0.48	0.14	0.75	−0.12	N/A[c]

[a]Performance of RF and *v*-NN consensus predictions. [b]Performance of RF predictions for compounds with contradictory *v*-NN predictions. [c]A total of 232 compounds had contradictory RF and *v*-NN predictions. [d]Performance of *v*-NN predictions for compounds with contradictory RF predictions. N/A: not applicable.

In summary, both the RF and the *v*-NN methods performed well when evaluated by 10-fold cross-validation with all the samples randomly distributed into 10 equal-sized groups. The consensus prediction given by the two methods was even more reliable. Furthermore, compounds with contradictory predictions by the two methods appear to be the ones for which neither method performed well.

For a QSAR model with this level of performance, one would expect the model to be predictive and practically useful. However, *n*-fold cross-validation with randomized samples is only one of the ways of evaluating model performance. Another and perhaps a more appropriate test of model quality is its performance for compounds that have not yet been synthesized/tested. As discovery research routinely explores new chemical space, prediction models developed from existing data and knowledge may not be appropriate for these novel chemical entities.

Performance of *De Novo* Predictions. To evaluate prediction performance of the *v*-NN and RF approaches based on time-dependent samples in the data, we grouped the HLM data by the year the data was published. In the data set, $T_{1/2}$ data of 1,738 compounds were reported from 1998 to 2010. In 2011, 2012, and 2013, $T_{1/2}$ data of 676, 461, and 516 additional compounds were reported, respectively. By the time we retrieved the HLM data from the ChEMBL database in February 2015, $T_{1/2}$ data of 226 compounds were associated with 2014 as the year of publication. In addition, 37 compounds in the data set were not associated with a publication date. Compared to the other years, the number of compounds published in 2014 appeared small. However, it is likely that data curation for 2014 was not complete yet and that only part of the data published in 2014 were available in the ChEMBL database.

To evaluate prospective prediction performance, we first used the data published prior to 2011 and data for the 37 compounds without a publication date as the training set for model development. We then used the resulting model to predict the rest of the data year by year as they were published from 2011 to 2014. We repeated this process using a stepwise transfer of the test-set data into the training set, i.e., increasing the size of the training set by adding data published in 2011, 2012, and 2013, successively, and predicting the rest of the data with the resulting models. For *v*-NN predictions, we used the parameters h = 0.20 and d_0 = 0.45, because they were the optimal parameters determined from the 10-fold cross-validation calculations. Model performance measures derived from these calculations are presented in Table 3 and graphically in Figure 3. They show, surprisingly, that neither of the methods withstood the test of time well. For *v*-NN, a major issue was the loss of coverage, with less than 10% of the compounds predictable. The RF method did not suffer a similar loss of coverage. However, its accuracy, sensitivity, and kappa

Table 3. Time-Dependent Performance of the *v*-NN and RF Models[a]

training set	method	accuracy	sensitivity	specificity	kappa	coverage
pre-2011 data	*v*-NN	0.68	0.24	0.82	0.06	0.08
	RF	0.55	0.15	0.81	−0.04	0.44
pre-2012 data	*v*-NN	0.71	0.25	0.95	0.30	0.08
	RF	0.61	0.28	0.80	0.09	0.51
pre-2013 data	*v*-NN	0.65	0.21	0.93	0.21	0.10
	RF	0.64	0.27	0.91	0.19	0.68
pre-2014 data	*v*-NN	0.81	1.00	0.50	0.62	0.07
	RF	0.66	0.25	0.93	0.21	0.64

[a]RF and *v*-NN models were built with data reported in the time periods in the training set column. We used the models to predict HLM stability of the rest of the compounds in the HLM data set, and the performance measures for these compounds are presented in this table.

DOI: 10.1021/acs.jcim.5b00255
J. Chem. Inf. Model. 2015, 55, 1566−1575

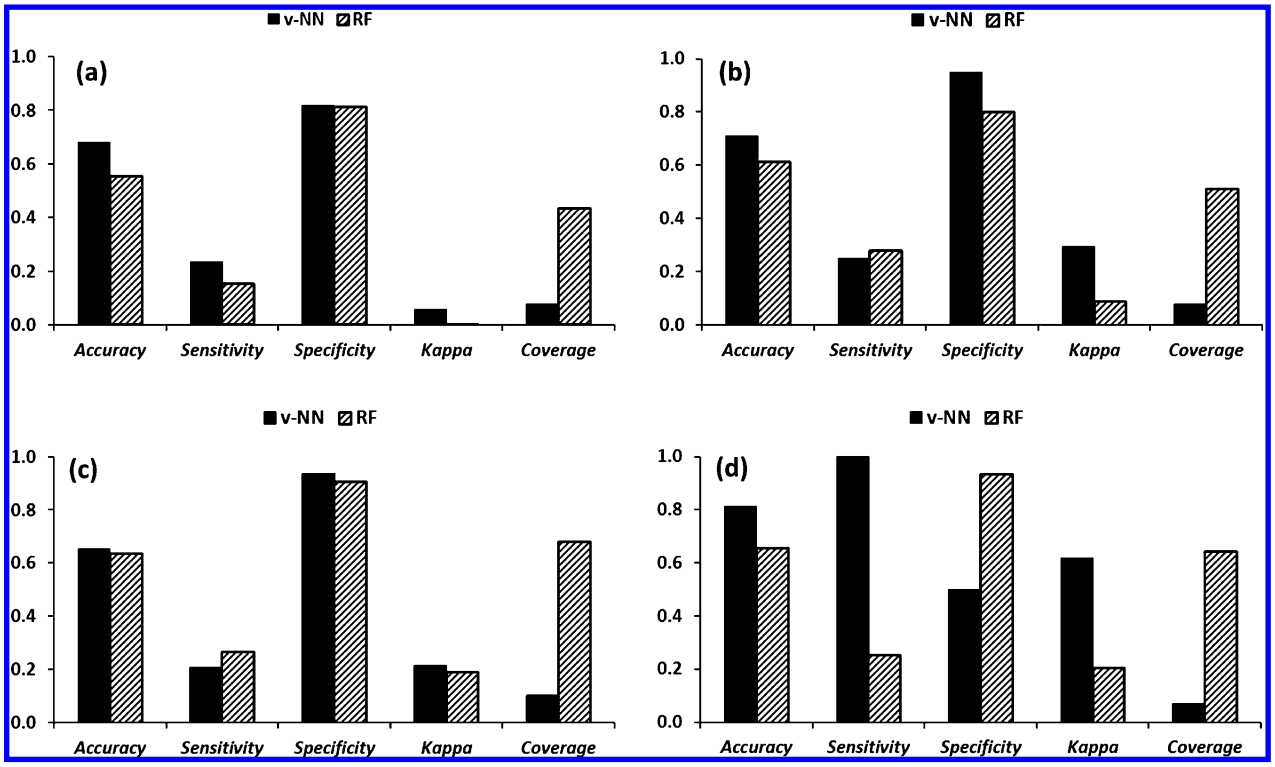

Figure 3. (a) Prediction performance for 2011−2014 data using a model built with pre-2011 data; (b) Prediction performance for 2012−2014 data using a model built with pre-2012 data; (c) Prediction performance for 2013−2014 data using a model built with pre-2013 data; (d) Prediction performance for 2014 data using a model built with pre-2014 data.

Table 4. Results of the 80:20 Split and y-Randomization Tests of the RF and *v*-NN Models Using HLM Data of 1998−2010[a]

evaluation method	model	accuracy	sensitivity	specificity	kappa	coverage
80:20 split tests	RF	0.84 (0.02)	0.79 (0.04)	0.87 (0.02)	0.66 (0.05)	0.87 (0.02)
	v-NN	0.83 (0.02)	0.75 (0.03)	0.88 (0.02)	0.64 (0.04)	0.92 (0.02)
y-randomization	RF	0.51 (0.03)	0.42 (0.06)	0.57 (0.04)	−0.02 (0.06)	0.87 (0.02)
tests	*v*-NN	0.54 (0.03)	0.31 (0.05)	0.69 (0.04)	0.00 (0.06)	0.92 (0.01)

[a]Numbers in parentheses are standard deviations of the performance measures calculated from 50 runs of the tests with samples randomly selected into the 80:20 training/test sets.

values were all significantly lower than the values estimated from the 10-fold cross-validation with time-randomized samples. Other than coverage, the performance measures of RF were slightly worse than those of *v*-NN. Both methods had very low sensitivity, rendering them practically unable to predict HLM-unstable compounds.

Model Evaluation with Independent Test Sets and y-Randomized Tests. Since the RF and *v*-NN models appeared highly predictive when evaluated by 10-fold cross-validation with time-randomized samples, but highly disappointing for prospective predictions, it raises the concern that the satisfactory performance exhibited in 10-fold cross-validation might be a result of overfitting. To assess this possibility, we performed calculations using the conventional 80:20 split of the pre-2011 data, building prediction models using 80% of the data and evaluate model performance with 20% of the data. As the performance measures may be variable depending on specific selection of the samples into the 20% test set, we repeated the test 50 times with the samples randomly segregated into the 80:20 split each time. In addition, we performed y-randomization tests. That is, with each 80:20 split, HLM stability classes of the training set compounds were randomly reassigned. The resulting data were then used to develop the RF and *v*-NN models. Model performance was evaluated using the test set. The y-randomized test was also repeated 50 times. The results are presented in Table 4 and compared with results of 10-fold cross-validation in Figure 4. They showed that when evaluated by the 80:20 split of the samples, the model performance is similar to that of 10-fold cross-validation. However, when the HLM stability classes of the training data were randomly reassigned, the resulting models have an overall rate of correct predictions of around 50%, similar to the chance from flipping a coin, and a kappa value close to zero−confirming that the chance of a correct prediction by the models was no better than a random guess. Thus, the y-randomization test confirmed that the RF and *v*-NN approaches captured molecular attributes that are important for molecular HLM stability, and the 80:20 split test confirmed that the satisfactory model performance as shown in 10-fold cross-validation was not due to overfitting.

Strategy To Mitigate Time-Dependent Performance Issues. Because both of the methods performed well for compounds whose chemistry space was adequately represented by the training set but showed significant underperformance for compounds not represented in the training set, a well-defined applicability domain is crucial for knowing when a prediction is

DOI: 10.1021/acs.jcim.5b00255
J. Chem. Inf. Model. 2015, 55, 1566−1575

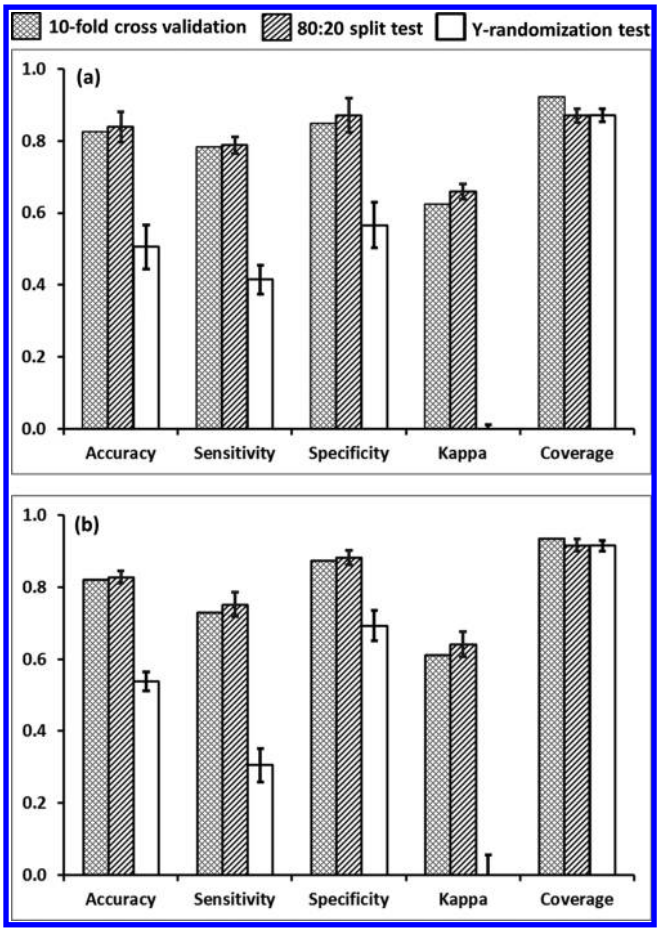

Figure 4. Results of the 80:20 split training/test evaluation and y-randomization tests compared to 10-fold cross-validation: (a) performance measures of RF models for the test set compounds, (b) performance measures of v-NN models for the test set compounds. The results are average of 50 runs with the training/test samples randomly selected. The error bars are ± one standard deviation.

reliable. More importantly, a well-defined applicability domain provides a path to expanding the applicability domain, as it identifies compounds outside the domain and draws attention to the need for experimental measurements of these compounds. Once experimental data from these compounds are included in the training set, the methods provide predictions for compounds in the previously underrepresented chemical space. To test this hypothesis, we retrained the models using the pre-2011 HLM data and a small percentage of randomly selected post-2010 data (1% to 20%) as the training set. We then made predictions for the rest of the post-2010 data and calculated the associated performance measures. Because the results varied with the randomly selected small percentage of post-2010 data included in the training set, we repeated the calculations 50 times and calculated the mean and the standard deviations of the performance measures. Table 5 and Figure 5 show that increasing the number of post-2010 compounds in the training set significantly improved the performance of both v-NN and RF for the post-2010 compounds. The improvements of both methods were mainly derived from the improved sensitivity and coverage, as the specificity remained essentially the same. The improvement in sensitivity led to significant improvement in the kappa values of both methods. For v-NN, including just 1% of post-2010 data (18 compounds) in the training set increased the sensitivity of v-NN for the other post-2010 compounds from 24% to 56%, kappa from a low 0.06 to 0.36, and coverage from 8% to 27%. The corresponding improvement in RF performance was not as significant, with an increase of sensitivity from 15% to 35%, kappa from −0.04 to 0.19, and coverage from 44% to 55%. When 20% of post-2010 compounds were included in the training set, both methods achieved adequate performance for the rest of the post-2010 compounds, with coverages of 74% (v-NN) and 81% (RF), kappa values of 0.46 (v-NN) and 0.44 (RF), specificities of 82% (v-NN) and 84% (RF), and sensitivities of 64% (v-NN) and 59% (RF).

The results in Table 5 and Figure 5 show that a practical strategy to expand the applicability domain of QSAR models is to retrain the models in a timely fashion with up-to-date experimental data in the training set. In this respect, v-NN is preferable to RF and many other machine-learning methods, as v-NN does not really build a physical model. Instead, it makes predictions on the fly simply by taking a weighted average of all available experimental data that meet a molecular structural similarity criterion. On the other hand, RF requires creation of a static mathematical model; therefore, when new experimental data become available, an existing model may become outdated, and a new static model may be needed to ensure maximum applicability.

Table 5. Time-Dependent Performance of the RF and v-NN Models with a Small Percentage of Test Data Included in the Training Set[a]

	accuracy		sensitivity		specificity		kappa		coverage	
% test data[b]	RF	v-NN	RF	v-NN	RF	v-NN	RF	v-NN	RF	v-NN
0	0.55	0.68	0.15	0.24	0.81	0.82	−0.04	0.06	0.44	0.08
1	0.64 (0.02)[c]	0.72 (0.04)	0.35 (0.06)	0.56 (0.11)	0.82 (0.04)	0.79 (0.07)	0.19 (0.06)	0.36 (0.09)	0.55 (0.03)	0.27 (0.04)
5	0.69 (0.03)	0.73 (0.03)	0.44 (0.07)	0.58 (0.08)	0.83 (0.04)	0.80 (0.04)	0.29 (0.07)	0.39 (0.07)	0.67 (0.03)	0.47 (0.03)
10	0.72 (0.02)	0.74 (0.02)	0.51 (0.05)	0.63 (0.05)	0.84 (0.03)	0.81 (0.04)	0.36 (0.04)	0.43 (0.04)	0.74 (0.03)	0.61 (0.02)
15	0.73 (0.01)	0.75 (0.02)	0.55 (0.04)	0.65 (0.04)	0.84 (0.03)	0.81 (0.03)	0.40 (0.03)	0.46 (0.03)	0.78 (0.03)	0.69 (0.02)
20	0.75 (0.01)	0.76 (0.02)	0.59 (0.04)	0.64 (0.04)	0.84 (0.02)	0.82 (0.02)	0.44 (0.03)	0.46 (0.04)	0.81 (0.03)	0.74 (0.02)

[a]The models were developed with pre-2011 data plus a small percentage of the post-2010 data as the training set; model performance for predicting the rest of the post-2010 data (test data) is presented in the table. [b]Percentage of randomly selected post-2010 data included in the training set for model development. [c]We performed 50 runs with randomly selected post-2010 data included in the training set. The average performance measures are presented; the standard deviations of the performance measures are given in parentheses.

DOI: 10.1021/acs.jcim.5b00255
J. Chem. Inf. Model. 2015, 55, 1566−1575

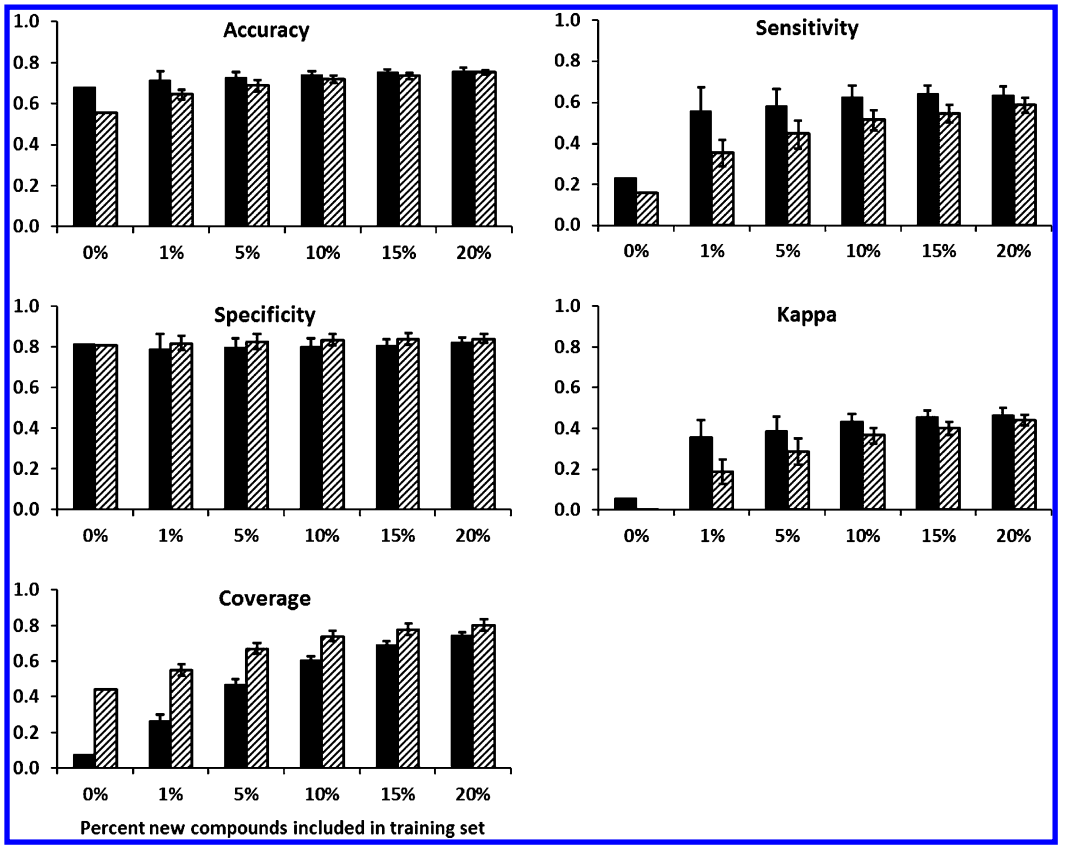

Figure 5. Prediction performance of *v*-NN (solid bar) and RF (striped bar) for 2011−2014 data using models developed with pre-2011 data and a small percentage of randomly selected post-2010 data included in the training sets. Bar heights are mean values of the performance measures over 50 runs; uncertainties are ± one standard deviation.

CONCLUSIONS

Niels Bohr was famously quoted as saying, "Prediction is very difficult, especially if it's about the future."[20] Indeed, the results of this study demonstrate that whereas both the *v*-NN and RF models can predict existing HLM data adequately, the models are of little value for prospectively predicting HLM stability for truly novel and unexplored chemical entities. This underscores the importance of assessing QSAR models in a time-dependent manner in order to be aware of the potential performance degradation associated with evaluating truly new chemistries. Our results fully support Sheridan's method of evaluating goodness of prospective predictions by time-split cross-validation.[12]

In addition, we also demonstrated that even though QSAR models developed with existing experimental data are unlikely to be capable of predicting compounds for future innovative discovery research, the value of QSAR modeling is not lost. Because retraining the models with the experimental data of a small number of new compounds can greatly enhance the applicability of the models. As an example, we demonstrated that with the *v*-NN and RF methods, including just 20% of new HLM data in the training set enabled the resulting models to predict 74% to 81% of the remaining new HLM data with ∼75% accuracy, a level of performance adequate for virtual screening. The key to success is to have a well-defined applicability domain in order to be able to readily identify compounds outside the domain, to perform experimental measurements on a few representatives of these compounds, and to immediately include the resulting data in the training set. In this respect, the *v*-NN method is superior, as it has a stringently defined applicability domain and makes predictions on the fly. As long as the *v*-NN method can access data from a repository of up-to-date experimental results, the *v*-NN predictions are always up to date.

ASSOCIATED CONTENT

Supporting Information

The HLM data set used in this study is provided in MS Excel format. It contains the ChEMBL IDs of the compounds, the SMILES strings, the HLM stability classes, and the 10-fold cross-validation groups to which the compounds belong. The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.5b00255.

AUTHOR INFORMATION

Corresponding Authors

*Phone: 301-619-1989. Fax: 301-619-1983. E-mail: RLiu@bhsai.org.

*E-mail: Sven.A.Wallqvist.civ@mail.mil.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors were supported by the U.S. Army Medical Research and Materiel Command (Fort Detrick, MD), as part of the U.S. Army's Network Science Initiative, and the Defense Threat Reduction Agency grant CBCall14-CBS-05-2-0007. The opinions and assertions contained herein are the private views of the authors and are not to be construed as official or as reflecting the views of the U.S. Army or of the U.S. Department

DOI: 10.1021/acs.jcim.5b00255
J. Chem. Inf. Model. 2015, 55, 1566−1575

of Defense. This paper has been approved for public release with unlimited distribution.

REFERENCES

(1) Mullard, A. New drugs cost US $2.6 billion to develop. *Nat. Rev. Drug Discovery* **2014**, *13* (12), 877−877.
(2) Fee, R. The cost of clinical trials. *Drug Discovery Dev.* **2007**, *10* (3), 32−32.
(3) Wang, J.; Urban, L. The impact of early ADME profiling on drug discovery and development strategy. *Drug Discovery World* **2004**, *Fall*, 73−86.
(4) Di, L.; Kerns, E. H.; Hong, Y.; Kleintop, T. A.; McConnell, O. J.; Huryn, D. M. Optimization of a higher throughput microsomal stability screening assay for profiling drug discovery candidates. *J. Biomol. Screening* **2003**, *8* (4), 453−462.
(5) Stepan, A. F.; Walker, D. P.; Bauman, J.; Price, D. A.; Baillie, T. A.; Kalgutkar, A. S.; Aleo, M. D. Structural alert/reactive metabolite concept as applied in medicinal chemistry to mitigate the risk of idiosyncratic drug toxicity: a perspective based on the critical examination of trends in the top 200 drugs marketed in the United States. *Chem. Res. Toxicol.* **2011**, *24* (9), 1345−410.
(6) Hu, Y.; Unwalla, R.; Denny, R. A.; Bikker, J.; Di, L.; Humblet, C. Development of QSAR models for microsomal stability: identification of good and bad structural features for rat, human and mouse microsomal stability. *J. Comput.-Aided Mol. Des.* **2010**, *24* (1), 23−35.
(7) Sakiyama, Y.; Yuki, H.; Moriya, T.; Hattori, K.; Suzuki, M.; Shimada, K.; Honma, T. Predicting human liver microsomal stability with machine learning techniques. *J. Mol. Graphics Modell.* **2008**, *26* (6), 907−915.
(8) Lee, P. H.; Cucurull-Sanchez, L.; Lu, J.; Du, Y. J. Development of in silico models for human liver microsomal stability. *J. Comput.-Aided Mol. Des.* **2007**, *21* (12), 665−73.
(9) Zakharov, A. V.; Peach, M. L.; Sitzmann, M.; Filippov, I. V.; McCartney, H. J.; Smith, L. H.; Pugliese, A.; Nicklaus, M. C. Computational tools and resources for metabolism-related property predictions. 2. Application to prediction of half-life time in human liver microsomes. *Future Med. Chem.* **2012**, *4* (15), 1933−44.
(10) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2012**, *40* (Database issue), D1100−D1107.
(11) Chen, B.; Sheridan, R. P.; Hornak, V.; Voigt, J. H. Comparison of random forest and Pipeline Pilot Naive Bayes in prospective QSAR predictions. *J. Chem. Inf. Model.* **2012**, *52* (3), 792−803.
(12) Sheridan, R. P. Time-split cross-validation as a method for estimating the goodness of prospective prediction. *J. Chem. Inf. Model.* **2013**, *53* (4), 783−90.
(13) 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* (3), 203−19.
(14) Obach, R. S.; Baxter, J. G.; Liston, T. E.; Silber, B. M.; Jones, B. C.; MacIntyre, F.; Rance, D. J.; Wastall, P. The prediction of human pharmacokinetic parameters from preclinical and in vitro metabolism data. *J. Pharmacol. Exp. Ther.* **1997**, *283* (1), 46−58.
(15) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50* (5), 742−54.
(16) Liu, R.; Tawa, G.; Wallqvist, A. Locally weighted learning methods for predicting dose-dependent toxicity with application to the human maximum recommended daily dose. *Chem. Res. Toxicol.* **2012**, *25* (10), 2216−26.
(17) Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5−32.
(18) 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. Model.* **2003**, *43* (6), 1947−58.
(19) R: The R project for Statistical Computing. http://www.r-project.org/ (accessed March 31, 2015).
(20) Bohr, N. BrainyQuote.com. http://www.brainyquote.com/quotes/quotes/n/nielsbohr130288.html (accessed March 1, 2015).
(21) Shen, M.; Xiao, Y.; Golbraikh, A.; Gombar, V. K.; Tropsha, A. Development and validation of k-nearest-neighbor QSPR models of metabolic stability of drug candidates. *J. Med. Chem.* **2003**, *46* (14), 3013−20.
(22) Dunn, G.; Everitt, B. *Clinical Biostatistics: An Introduction to Evidence-based Medicine*; Edward Arnold: London, 1995.
(23) Liu, R.; Wallqvist, A. Merging applicability domains for in silico assessment of chemical mutagenicity. *J. Chem. Inf. Model.* **2014**, *54* (3), 793−800.
(24) Rydberg, P.; Gloriam, D. E.; Olsen, L. The SMARTCyp cytochrome P450 metabolism prediction server. *Bioinformatics* **2010**, *26* (23), 2988−9.

DOI: 10.1021/acs.jcim.5b00255
J. Chem. Inf. Model. 2015, 55, 1566−1575