Journal of
Chemical Information
and Modeling

Article

pubs.acs.org/jcim

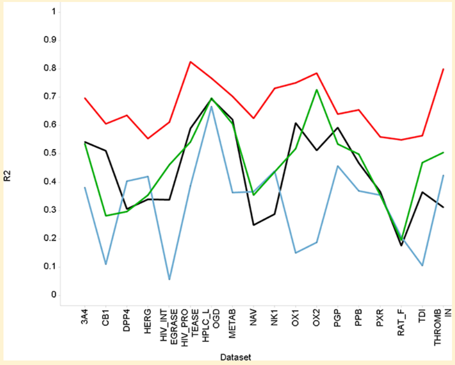
Time-Split Cross-Validation as a Method for Estimating the Goodness of Prospective Prediction.

Robert P. Sheridan*

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

Supporting Information

ABSTRACT: Cross-validation is a common method to validate a QSAR model. In cross-validation, some compounds are held out as a test set, while the remaining compounds form a training set. A model is built from the training set, and the test set compounds are predicted on that model. The agreement of the predicted and observed activity values of the test set (measured by, say, R^2) is an estimate of the self-consistency of the model and is sometimes taken as an indication of the predictivity of the model. This estimate of predictivity can be optimistic or pessimistic compared to true prospective prediction, depending how compounds in the test set are selected. Here, we show that time-split selection gives an R^2 that is more like that of true prospective prediction than the R^2 from random selection (too optimistic) or from our analog of leave-class-out selection (too pessimistic). Time-split selection should be used in addition to random selection as a standard for cross-validation in QSAR model building.



INTRODUCTION

There are a number of approaches for judging how well a QSAR model makes predictions. The most common is cross-validation. Here, we are using the term "cross-validation" in the sense that one holds out a subset of the data as the test set, makes a QSAR model on what is left (the training set), predicts the activities of the test set, and compares the predictions to the observed activities in the test set. (This is in contrast to getting the test set from a different source than the training set.) It is well understood that, at best, cross-validation can give a measure of the self-consistency of the data,[1] but strictly speaking cannot give a true estimate of the predictability of a model because one does not necessarily know in advance that the compounds being predicted will be in the "domain of the model". The field of "domain applicability"[2−17] is meant to determine which molecules can be reliably predicted on a specific model and which cannot.

In the pharmaceutical environment, a model is made from compounds tested in the appropriate assay at the time the model was made and is used to predict the activity of compounds not yet tested. Some of the new compounds may be analogs of compounds in the training set and some may not. Practitioners of QSAR are often asked by medicinal chemists how accurate predictions from a QSAR model will be on compounds not yet synthesized, and the chemists rarely accept "Depends on whether the individual compound is in the domain of the model" as an answer. Generally, they are interested in the overall accuracy of all predictions, not just a subset of compounds for which the prediction is deemed reliable. Here, we try three types of cross-validation to quantitatively estimate the prospective predictivity for a QSAR model in real-world situations.

The type of cross-validation most often used in the literature is based on random selection. That is, some fraction of compounds is randomly extracted from the data set to form a test set. One alternative suggestion is to construct training and test sets from compounds from the same clusters (i.e., "rational selection").[18−20] One would expect rational selection to produce test sets where most molecules in the test set would have an analog in the training set. Random selection would approximate this if the data set were large enough, so there is no particular advantage for rational selection over random selection. This has been recently demonstrated by Martin et al.[19] At the other extreme, the leave-class-out[21] method ensures that the molecules in the test set are not analogs of compounds in the training set. One would expect random selection and rational selection to produce optimistic estimates of prediction, while leave-class-out would be pessimistic. In an earlier paper,[22] we suggested "time-split" as a useful type of cross-validation, i.e., one builds a model on assay data available at a certain date and tests the model on data that is generated later, in effect more closely simulating the process of prospective validation. In this paper, we demonstrate that the R^2 for time-split cross-validation is much closer to the R^2 of prospective prediction than the R^2 of random-split or our equivalent of leave-class-out (neighbor-split). We can also show that random-split validation and neighbor-split validation are optimistic and pessimistic, respectively, mostly because of the similarity of the nearest compound in the training set.

Received: February 4, 2013
Published: March 22, 2013

dx.doi.org/10.1021/ci400084k | J. Chem. Inf. Model. 2013, 53, 783−790

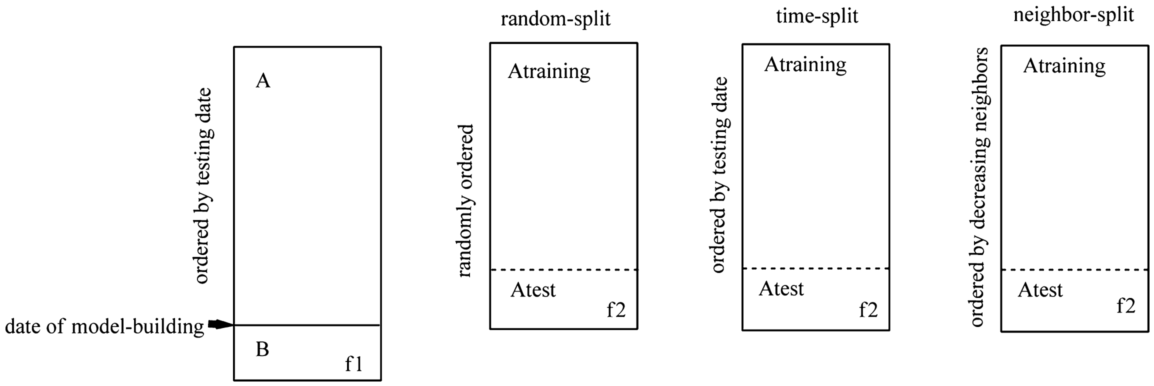


Figure 1. Scheme for cross-validations. B represents molecules for true prospective predictions. A represents the molecules that will be partitioned into training and test sets in various ways. The goal is to see which method of partitioning will give the goodness of prediction of A_{test} closest to that of B.

■ METHOD

We are following the scheme in Figure 1. To simulate prospective prediction on a data set of N_{total} compounds, we take the last f_1N_{total} compounds as set B. The remaining compounds (set A) are used to build a model. The R^2 of predictions of compounds in B against the model from A will be our standard of prospective prediction.

Descriptors and QSAR Method. All models are built as regressions using the random forest method[23,24] and the union of AP and DP descriptors, which in our hands gives the most accurate predictions. AP is the original Carhart "atom pair" descriptor,[25] and DP is the physiochemical analog (called BP in Kearsley et al.[26]). Both descriptors are of the form:

Atom type i − (distance in bonds) − Atom type j

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

Here, we are using 50 random forest trees for each model. The accuracy of prediction plateaus at ∼50 trees, so using more trees has no advantage.

Cross-Validation Schemes. Set A will be cross-validated in three ways. Let N_A be the number of compounds in A. In each case a fraction f_2 of the compounds in A are chosen to form the test set A_{test}, and the remaining compounds form $A_{training}$. The R^2 for the prediction of A_{test} on the model built from $A_{training}$ is generated. The idea is to see which flavor of A_{test} produces an R^2 closet to the R^2 of set B.

1. Random-split. f_2N_A compounds are randomly assigned as A_{test}. Normally in the literature there would be multiple random-splits; however, because the other cross-validation schemes involve only a single split, we do that here. If the data set is large enough, we would not expect different random splits to behave very differently.
2. Time-split. The last tested f_2N_A compounds in A are assigned as A_{test}.
3. Neighbor-split. The f_2N_A compounds in A with the fewest neighbors are assigned as A_{test}. Compounds are "neighbors" if their similarity ≥ 0.7 based on the AP descriptor and the Dice similarity index. This is meant to mimic certain aspects of "leave-class-out," where compounds are tested against a training set where there are no close analogs of the compounds. In true leave-class-out, all compounds in a given cluster would be left out at one time, and these clusters would be of different sizes, whereas here A_{test} has to be the same size for all cross-validation schemes.

Here, we set f_1 as 0.1 and f_2 as 0.1, 0.25, and 0.5. The value 0.1 represents the situation where almost all of the compounds are in the training set. The value 0.25 is the fraction for time-split we used in our previous paper.[22] The value 0.5 represents the largest fraction for cross-validation usually seen in the literature.

Domain Applicability Metrics. While we are not doing explicit domain applicability calculations in the sense that we are eliminating compounds when calculating R^2, here we can use domain applicability metrics as a way to explain why the different cross-validation schemes have systematically different R^2. The following are the metrics we investigated for B relative to A and the same apply to A_{test} relative to $A_{training}$:

1. For each compound in B, find the number of neighbors in A (NNEIGHBORS). The definition of neighbor is the same as the one above. The metric is the mean number of neighbors over all compounds in B. Previously, we found that the number of neighbors is a reasonable domain applicability metric: more neighbors indicates more accurate predictions.[2]
2. For each compound in B find the similarity (AP/Dice) to the nearest compound in A (SIMILARITYNEAREST1). The metric is the mean SIMILARITYNEAREST1 over all compounds in B. SIMILARITYNEAREST1 is also recognized as a reasonable domain applicability metric: higher SIMILARITYNEAREST1 means more accurate predictions.[2]
3. We have shown[3] that the variation of predicted value among random forest trees (TREESD) is a very discriminating domain applicability metric (lower TREESD means more accurate predictions). TREESD is a more discriminating metric than NNEIGHBORS and SIMILARITYNEAREST1. For each compound in B, find the TREESD. Because each data set has a different range of activity, this must be normalized:

   normTREESD = TREESD/STDEV observed activity.

   Our metric is the mean normTREESD over all compounds in B.

Data Sets. The data sets we use here are in Table 1. They are meant to be a representative mixture of fairly large pharmaceutically relevant data set, some on-target, some ADME related. Some are diverse, some less so. These data sets have been used in a previous publication.[22] In this publication, we are using the

Table 1. Data Sets for Prospective Prediction

Data set	Description	N_{total}
3A4	CYP 3A4 inhibition -log(IC50) M	50000
CB1[1]	CB1 binding -log(IC50) M	11640
DPP4[1]	DPP4 inhibition -log(IC50) M	8327
HERG	HERG inhibition -log(IC50) M	50000
HIV_INTEGRASE[1]	HIV integrase cell based assay -log(IC50) M	2421
HIV_PROTEASE[1]	HIV protease inhibition -log(IC50) M	4311
HPLC_LOGD	logD measured by HPLC method	50000
METAB	percent remaining after 30 min microsomal incubation	2092
NAV	NAV1.5 inhibition -log(IC50) M	46245
NK1[1]	NK1 (substance P) receptor binding -log(IC50) M	13482
OX1[1]	Orexin 1 inhibition -log(KI) M	7135
OX2[1]	Orexin 2 inhibition -log(KI) M	14875
PGP	log(BA/AB) 1uM human	8603
PPB	human plasma protein binding log(bound/unbound)	11622
PXR	pregnane X receptor maximum activation (percent) relative to rifampicin	50000
RAT_F	log(rat bioavailability) at 2 mg/kg	7821
TDI	time dependent 3A4 inhibitions log(IC50 without NADPH/ IC50 with NADPH)	5559
THROMBIN[1]	human thrombin inhibition -log(IC50) M	6924

[1] On-target data set.

real-number activities instead of categories, i.e., our models are regressions rather than classifications.

RESULTS

Comparing Cross-Validation Methods. R^2 for B and the three cross-validation methods for the A_{test} are in the Supporting Information. Figure 2 shows the R^2 vs the data sets for $f_2 = 0.1$. B, which represents our "standard of truth" for prospective prediction, is represented by the black line. While A_{test} for random-split (red line) shows the same trend as B, its R^2 is much above that of B, and the range of R^2 is compressed. It is consistently too optimistic. A_{test} for neighbor-split (blue line) is not always lower than B, but in most cases it is, i.e., neighbor-split tends to be pessimistic. A_{test} for time-split (green) is closest to being superimposed on B (black). An alternative way of looking at the same data, such that the differences from the R^2 of B are more easily perceived, is to plot diff_R2 (R^2 for A_{test} minus R^2 for B). This is in Figure. 3. The plot of R^2 for A_{test} vs R^2 for B is shown at the top of Figure 4. Clearly time-split (green) fall closest to the diagonal, as we would expect from Figure 2. The plot of diff_R^2 for A_{test} vs R^2 for B is shown at the bottom of Figure 4. In the latter plot, we can more clearly see that the magnitude of optimism of and pessimism depends on the R^2 for B. The clearest relationship is that random-split is most over-optimistic (diff_$R^2 > 0$) when R^2 for B is low (R^2 for the linear fit = 0.64). A less clear trend (R^2 for the linear fit = 0.37) is that neighbor-split is most overpessimistic (diff_$R^2 < 0$) when R^2 for B is high. The trend for time-split is least compelling (R^2 for the linear fit = 0.20) and has a more horizontal slope relative to the other two, again consistent with the observation in Figure 2 that time-split matches B fairly well over the range of R^2 compared to the other cross-validation schemes.

Effect of f_2. Given that $f_1 = 0.1$, does f_2 also need to be 0.1 to make a good match? Figure 5 shows diff_R^2 as a function of f_2. Generally, there does not seem to be an overall trend for random-split and time-split; the red and green lines are more or less horizontal. The only visible trend is for diff_R^2 of neighbor-split (blue lines) to rise, i.e., for neighbor-split to be less pessimistic, (especially for OX1 and CB1) as f_2 increases. This is not unexpected, as f_2 gets larger, there will be proportionately fewer compounds in A_{test} with few neighbors in A_{train}, and predictions should get better. This implies that a reasonable estimate of R^2 for B can be gotten from time-split using any value of f_2 in the range of 0.1 to 0.5.

Domain Applicability Metrics. One expects the goodness of prediction to track with domain applicability metrics. However, which metric will prove more important? That is, does random-split overestimate R^2 because NNEIGHBORS in A_{test} is too high relative to what we see in B, SIMILARITYNEAREST1is too high, or TREESD is too low? Figures 6 through 8 show diff_R^2 vs the difference in domain applicability metrics (A_{test} minus B) for $f_2 = 0.1$. Very similar plots are seen with $f_2 = 0.25$ and 0.5. In all cases the time-split method (green circles) is near diff_$R^2 = 0$. That is, time-split approximates B for all of the metrics, whereas random-split (red) is too optimistic (higher mean NNEIGHBORS, higher SIMILARITYNEAREST1, and lower meanTREESD) and neighbor-split (blue) is too pessimistic. Interestingly, SIMILARITYNEAREST1 appears to show the clearest (albeit far from perfect) trend, with the best correlation with diff_R^2 (R^2 for the linear fit = 0.58) and the clearest separation of time-split from the other methods. This is unexpected because previously we found that TREESD is a much more discriminating parameter for domain applicability than SIMILARITYNEAREST1[3]. At present, we have no explanation. Note that although SIMILARITYNEAREST1 explains the differences between the cross-validation schemes fairly well, i.e., the red, green, and blue clusters in Figure 7 are well separated and form a straight line, it does not much explain the differences between data sets within each scheme, i.e., the trends within the red, green, and blue clusters are weaker.

DISCUSSION

Merck recently participated in a Kaggle competition (http://www.kaggle.com/c/MerckActivity) to test our in-house QSAR methodology against other methods in the machine-learning community. We used a subset of the data sets discussed here and constructed training and validation sets by time-split. We were surprised how many Kaggle contestants were puzzled by the fact that the distributions of descriptors in sets A and B (in the nomenclature of this paper) were different; the expectation seemed to be that A and B would be randomly selected from the same pool. A number of machine-learning methods have adjustable parameters that investigators typically calibrate using random selections from set A, and the fact that the descriptor distributions in B were different would at least partly frustrate

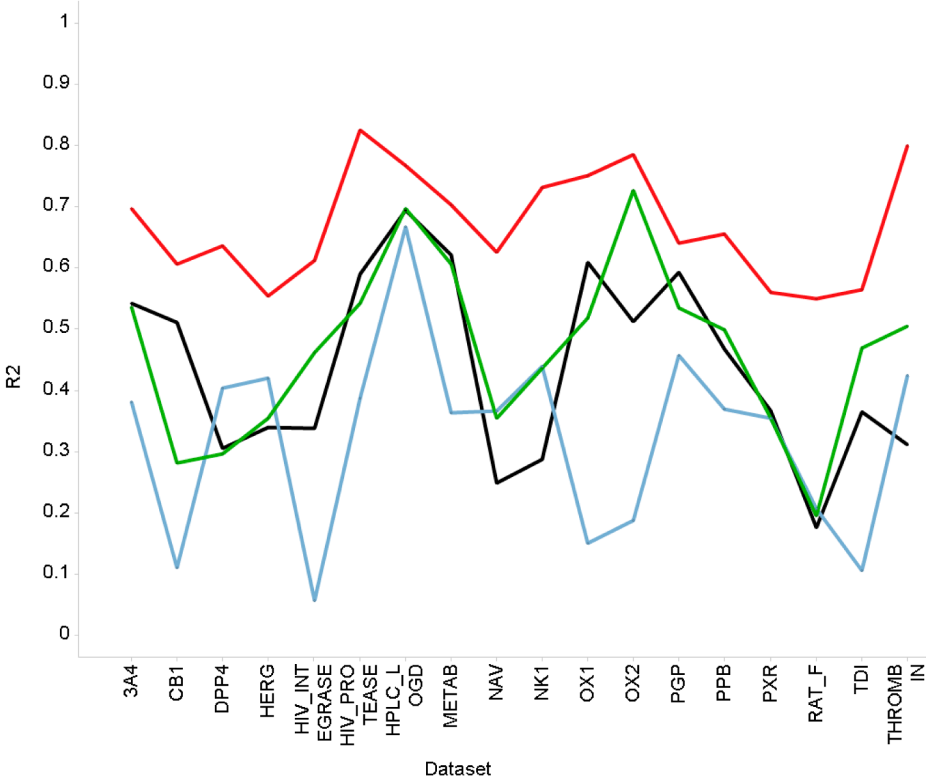


Figure 2. R^2 for prospective prediction (black) and the three cross-validation schemes for the data sets where $f_2 = 0.1$: random-split (red), time-split (green), neighbor-split (blue).

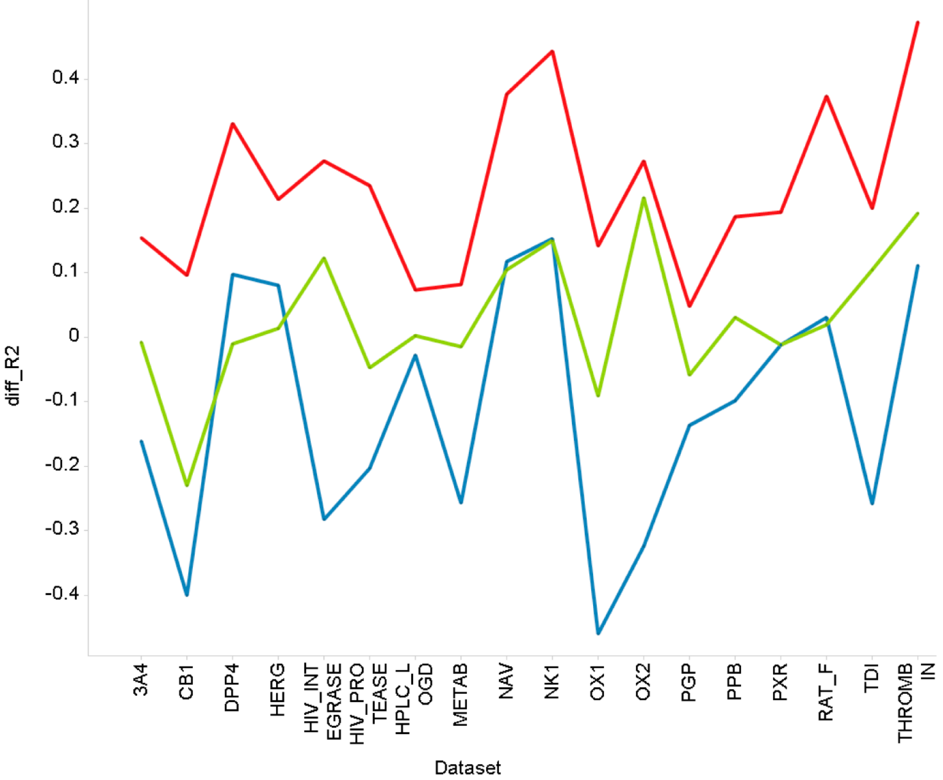


Figure 3. Difference in R^2 (A_{test} minus B) for the data sets where $f_2 = 0.1$: random-split(red), time-split (green), and neighbor-split (blue).

that approach. (Adjustable parameters are not so much an issue with random forest.) This did show us that time-split validation is not very familiar in the machine-learning community, and that fact encouraged us to publish this work. Time-split is implicitly done by a number of workers in studies monitoring how fast model accuracy falls off after a model has been built.[27,28]

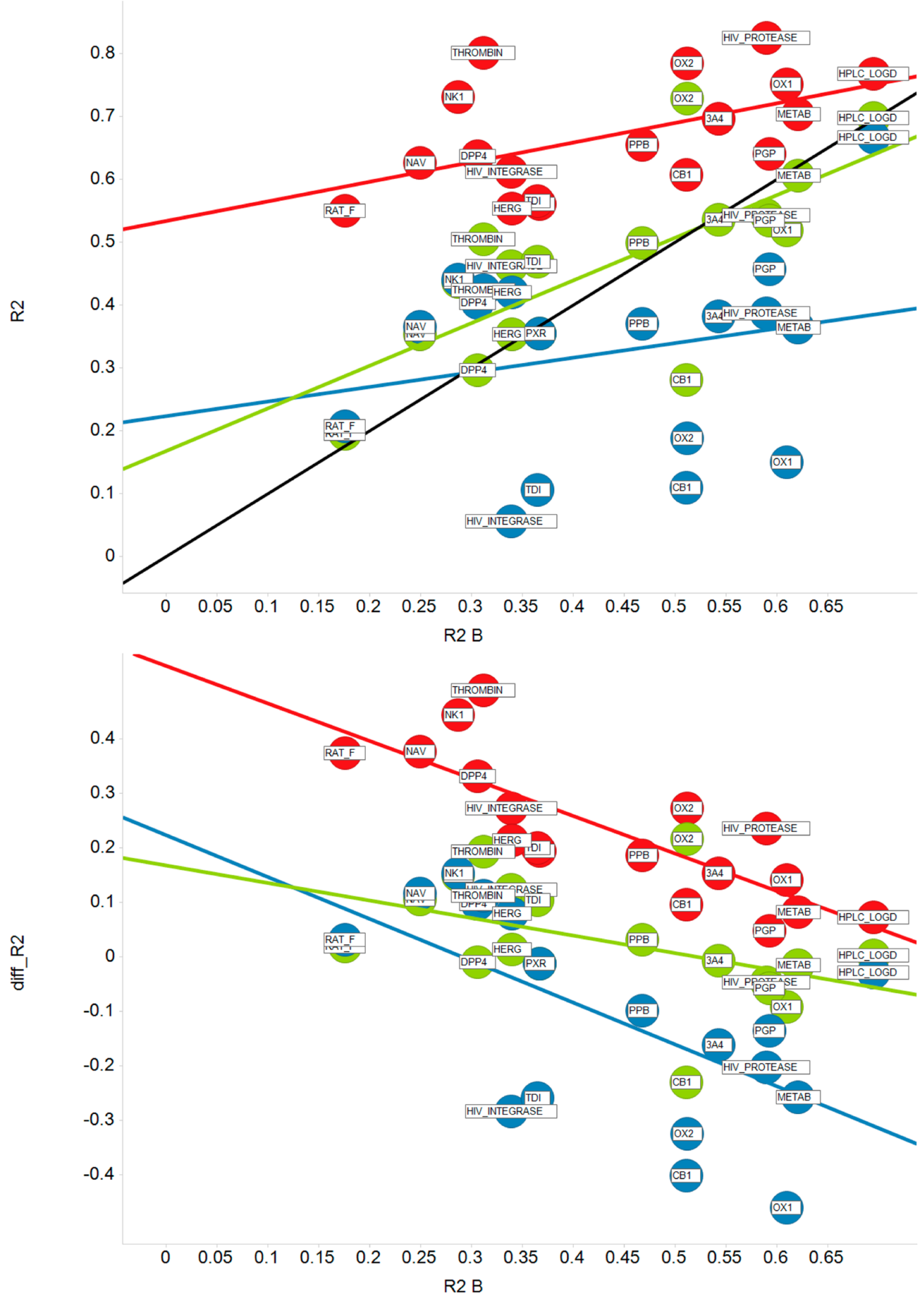


Figure 4. Top: R^2 for cross-validation vs R^2 for B. Black line is the diagonal. Bottom: Difference in R^2 vs R^2 for B for $f_2 = 0.1$. Each circle represents a data set: random-split (red), time-split (green), and neighbor-split (blue).

During the revision of this paper, we became aware of an accepted, but not yet published, paper by Wood et al.,[29] which also noted that random-split cross-validation greatly overestimates the prediction accuracy relative to prospective prediction.

Why does time-split cross-validation approximate true prospective prediction? Given that the difference in mean SIMILARITYNEAREST1 shows a useful trend, we can imagine a situation where compounds to be predicted prospectively are all close analogs of some compound in the model, in which case the R^2 from the random-split cross-validation would be a reasonable approximation of the R^2 of B. We can also imagine a situation where, after the model was built, there was an extreme change in compound class, in which case R^2 from neighbor-split would most closely approximate the R^2 from B. In most medicinal chemistry programs in progress, the true situation is somewhere in the middle. New chemical classes are added with time, and old ones are dropped, but some chemical classes continue. This happens regularly enough that, in the limit of large data sets, random-split is on the average too optimistic and neighbor-split too pessimistic. Time-split, by simulating prospective predictions within set A, can better approximate true prospective prediction, assuming the rate of change in chemical series is comparable before and after the model is built.

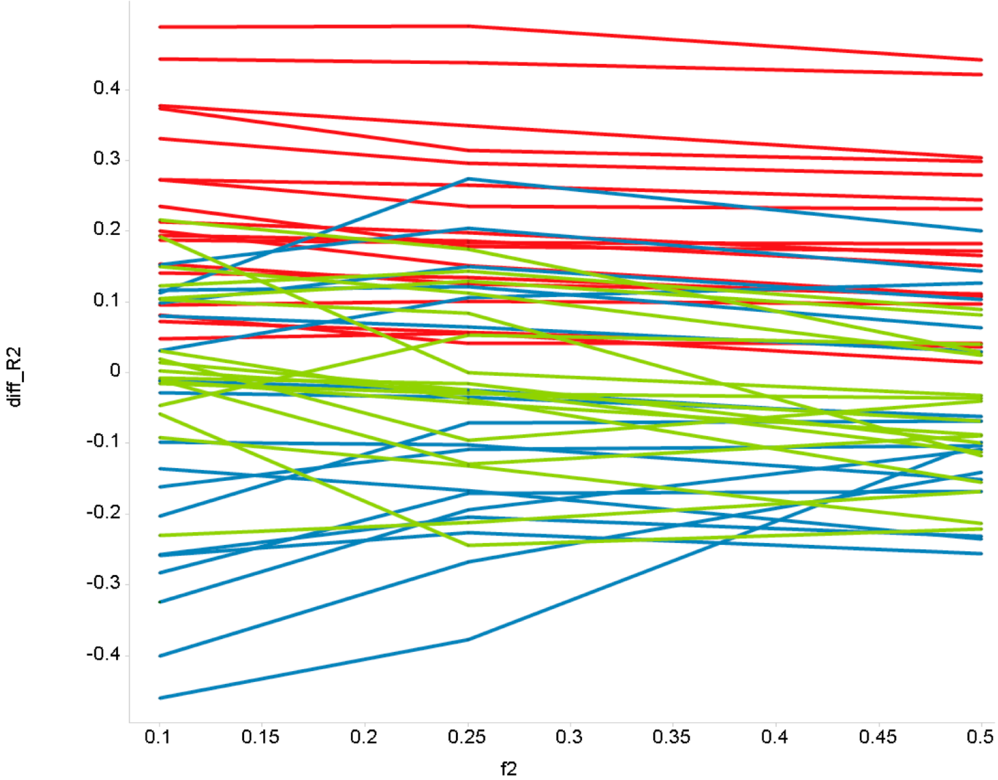


Figure 5. Difference in R^2 vs f_2. Each line represents a data set: random-split(red), time-split (green), and neighbor-split (blue).

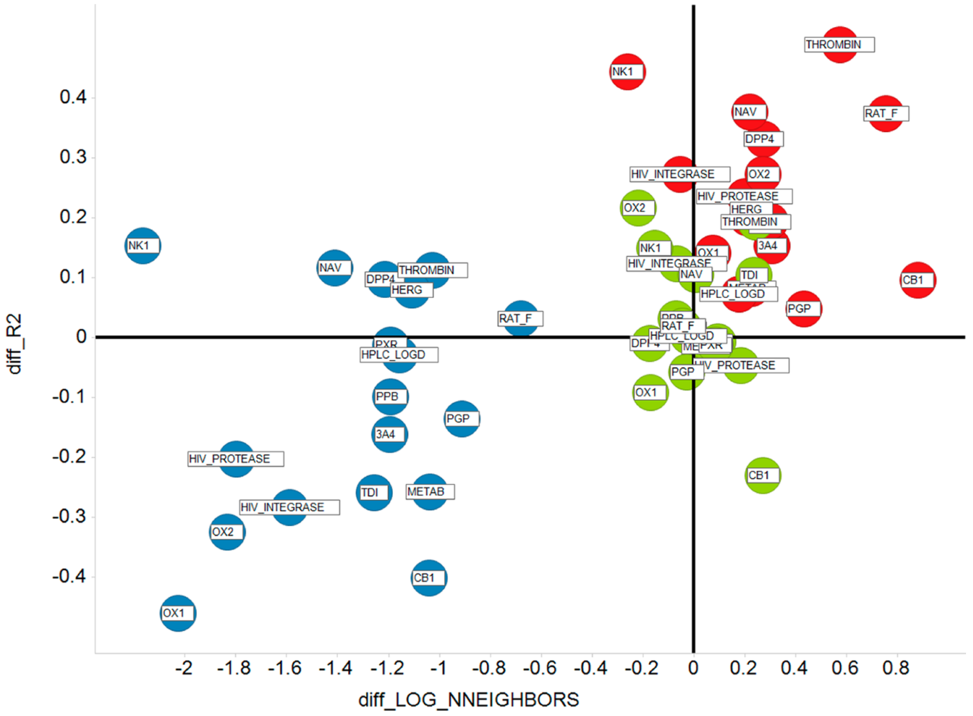


Figure 6. Diff_R^2 vs the difference in log_meanNNEIGHBORS for $f_2 = 0.1$. Each circle represents a data set: random-split(red), time-split (green), and neighbor-split (blue).

Given that it is likely to be too optimistic, why is random-split the standard type of cross-validation in QSAR and other machine-learning fields? One reason is that it is trivial to execute. Another, and probably the most important, is that one does not need to know dates of testing, which are at present seldom available for any data except that produced in one's home institution. It would be very useful to be able to predict the prospective R^2 directly from a cross-validated R^2 using random-split. However, as shown in the top of Figure 4, the correlation between R^2 B and R^2 for A_{test} random-split is too low to allow this. We encourage those who publish data sets to include dates of testing so time-split validation is possible.

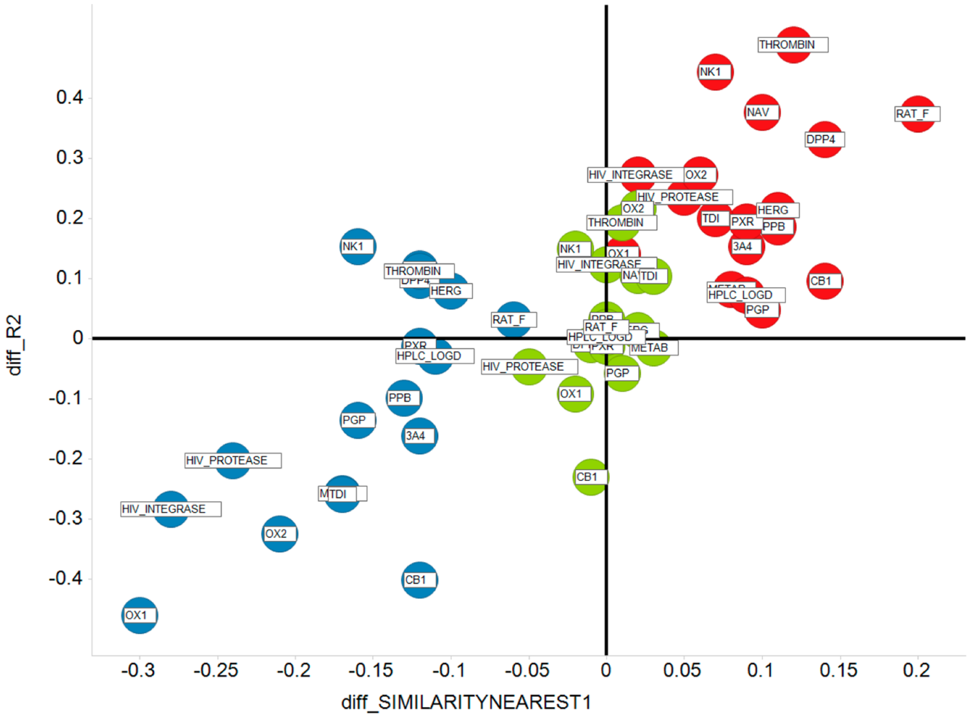


Figure 7. Diff_R^2 vs difference in meanSIMILARITYNEAREST1 for f_2 = 0.1. Each circle represents a data set: random-split(red), time-split (green), and neighbor-split (blue).

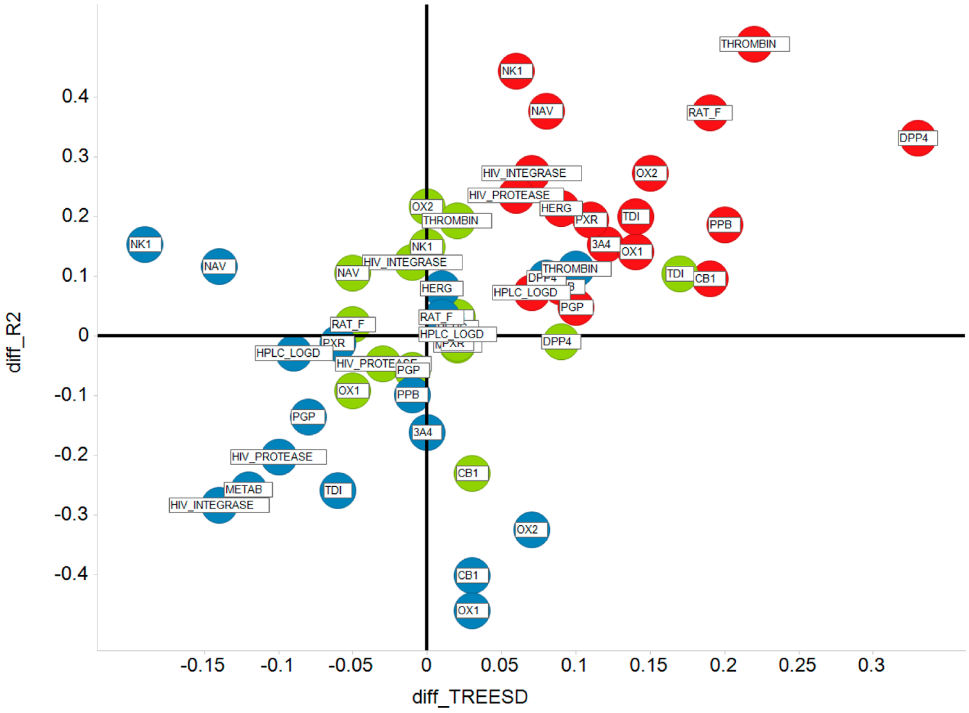


Figure 8. Diff_R^2 vs difference in mean-normTREESD for f_2 = 0.1. Each circle represents a data set: random-split(red), time-split (green), and neighbor-split (blue).

■ ASSOCIATED CONTENT

Supporting Information

Table of R^2 and domain applicability metrics for all data sets at three values of f_2. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: sheridan@merck.com.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

Joe Shpungin wrote the parallelized version of random forest that we use here. The In Silico ADME working group first suggested time-split as a reasonable method of estimating the accuracy of true prospective prediction. A large number of Merck biologists, over many years, generated the data for examples used in the paper.

■ REFERENCES

(1) Golbraikh, A.; Tropsha, A. Beware of q2. *J. Mol. Graphics Modell.* **2002**, *20*, 269−276.

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

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

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

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

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

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

(8) He, L.; Jurs, P. C. Assessing the reliability of a QSAR model's predictions. *J. Mol. Graphics Modell.* **2005**, *23*, 503−523.

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

(10) Schroeter, T. S.; Schwaighofer, A.; Mika, S.; Laak, A. T.; Suelzle, D.; Ganzer, U.; Heinrich, N.; Muller, K.-R. Estimating the domain of applicability for machine learning QSAR models: A study on aqueous solubility of drug discovery molecules. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 651−664.

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

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

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

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

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

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

(17) Sahlin, U.; Filipsson, M.; Öberg, T. A risk assessment perspective of current practice in characterizing uncertainties in QSAR regression predictions. *Mol. Inf.* **2011**, *30*, 551−564.

(18) Golbraikh, A.; Tropsha, A. Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 357−369.

(19) Martin, T. M.; Harten, P.; Young, D. M.; Muratov, E. N.; Golbraikh, A.; Zhu, H.; Tropsha, A. Does rational selection of training and test sets improve the outcome of QSAR modeling. *J. Chem. Inf. Model.* **2012**, *52*, 2570−2578.

(20) Leonard, J. T.; Roy, K. On selection of training and test sets for the development of predictive QSAR models. *QSAR Comb. Sci.* **2006**, *25*, 235−251.

(21) Lombardo, F.; Obach, R. S.; Shalaeva, M. Y.; Gao, F. Prediction of human volume of distribution values for neutral and basic drugs. 2. Extended data seta and leave-class-out statistics. *J. Med. Chem.* **2004**, *47*, 1242−1250.

(22) Chen, B.; Sheridan, R. P.; Hornak, V.; Voigt, J. H. Comparison of random forest and Pipeline Pilot Naïve Bayes in prospective QSAR predictions. *J. Chem. Inf. Model.* **2012**, *52*, 792−803.

(23) Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 3−32.

(24) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. Random forest: a classification and regression tool for compound classification and QSAR modeling. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1947−1958.

(25) Carhart, R. E.; Smith, D. H.; Ventkataraghavan, R. Atom pairs as molecular features in structure-activity studies: Definition and application. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64−73.

(26) Kearsley, S. K.; Sallamack, S.; Fluder, E. M.; Andose, J. D.; Mosley, R. T.; Sheridan, R. P. Chemical similarity using physiochemical property descriptors. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 118−27.

(27) Rodgers, S. L.; Davis, A. M.; Tomkinson, N. P.; van de Waterbeemd, H. Predictivity of simulated ADME autoQSAR models over time. *Mol. Inf* **2011**, *30*, 256−266.

(28) Wood, D. J.; Buttar, D.; Cumming, J. G.; Davis, A. M.; Norinder, U.; Rodgers, S. L. Automated QSAR with a hierarchy of global and local models. *Mol. Inf.* **2011**, *30*, 960−972.

(29) 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**, DOI: 10.1007/s10822-013-9639-5.