JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

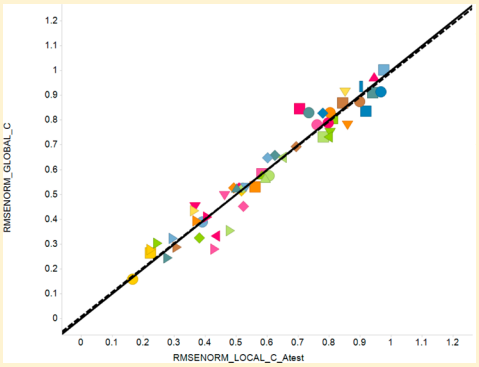
Global Quantitative Structure−Activity Relationship Models vs Selected Local Models as Predictors of Off-Target Activities for Project Compounds

Robert P. Sheridan*

Cheminformatics Department, RY800-D133, Merck Research Laboratories, Rahway, New Jersey 07065, United States

Supporting Information

ABSTRACT: In the pharmaceutical industry, it is common for large numbers of compounds to be tested for off-target activities. Given a compound synthesized for an on-target project *P*, what is the best way to predict its off-target activity *X*? Is it better to use a global quantitative structure−activity relationship (QSAR) model calibrated against all compounds tested for *X*, or is it better to use a local model for *X* calibrated against only the set of compounds in project P? The literature is not consistent on this topic, and strong claims have been made for either. One particular idea is that local models will be superior to global models in prospective prediction if one generates many local models and chooses the type of local model that best predicts recent data. We tested this idea via simulated prospective prediction using in-house data involving compounds in 11 projects tested for 9 off-target activities. In our hands, the local model that best predicts the recent past is seldom the local model that is best at predicting the immediate future. Also, the local model that best predicts the recent past is not systematically better than the global model. This means the complexity of having project- or series-specific models for *X* can be avoided; a single global model for *X* is sufficient. We suggest that the relative predictivity of global vs local models may depend on the type of chemical descriptor used. Finally, we speculate why, contrary to observation, intuition suggests local models should be superior to global models.



INTRODUCTION

In the pharmaceutical industry it is common for large numbers of compounds to be tested for off-target activities, which need to be taken into account in the development of a therapeutic program. Given a compound *M* in a project *P* aimed at a particular target (for example, HIV integrase), what is the best type of quantitative structure−activity relationship (QSAR) model to predict its off-target activity *X* (for example, binding to the HERG channel)? One can imagine three types of models, in order of increased specificity:

1. a "global" model generated from all compounds tested for *X*.
2. a "local" (also called "project") model generated from only the set of compounds made for project *P* that are tested for *X*.
3. a "series" model generated from a set of analogs of *M* in project *P* that are tested for *X*.

Note that QSAR models for on-target activities, i.e. the target receptor for *P*, will always be local or series.

Obviously, it would be simpler to have a single global model for *X* to cover all projects and series. However, intuition suggests that local models will be more accurate because they are calibrated only on the relevant chemical series (or a single series) for each *P*. There is also the consideration that in practice it might be easier to update many smaller models than one large model. The topic of global vs local has been debated in the recent literature; for example, refs 1−9. Helgee et al.[2] did a study wherein they claimed that global models were at least as good or better than local models, and this was especially true for models made with the random forest method. Two limits of that study were that test sets were generated by random selection of entire data sets—whereas, a more realistic test would have involved prospective prediction—and that the local models involved only those molecules most similar to the compound being predicted, as in "lazy learning".[6−9] The local models of Helgee et al. could thus be considered more like series models. Other workers[3−5] describe systems wherein they can build local models superior to global models. One that we recently found most interesting is that of Wood et al.[3] because it used simulated prospective prediction.

Wood et al.[3] looked at three physicochemical properties (logD, plasma protein binding, and solubility) and nine (unnamed) projects and series therein. One interesting type of plot in the works of Wood et al.[3] and Davis and Wood[5] is the root-mean-square error (RMSE) of prediction for one assay *X* (averaged over many projects) as a function of time after a model is built, usually in increments of 1 or 2 months. The variation in RMSE month to month is large, but generally there

Received: February 11, 2014
Published: March 14, 2014

   dx.doi.org/10.1021/ci500084w | *J. Chem. Inf. Model.* 2014, 54, 1083−1092

is an overall increase in RMSE with time, i.e. the models become less predictive, although this trend is not clear until many months have passed. The relative utility of different types of model generation can be compared on these plots. A recently updated global model will clearly beat a global model that is static (i.e., not updated) at prospective prediction. This is not at all surprising since new molecules are increasingly outside the scope of a static model. Wood et al.[3] also concluded that a "hierarchy" of models will beat an updated global model at prospective prediction, although this effect is not as strong.

The hierarchical approach suggested by Wood et al.[3] involves choosing the best model for compounds in *P* out of a set of global, local, and series models made using various QSAR methods (random forest only for global models; random forest and partial least squares (PLS) for local and series models). The "best" model is the data set/QSAR method combination that gives the most accurate retrospective predictions on a small set of recently tested molecules. Presumably the same combination will give the best prospective prediction of new molecules. This is the basis of the AutoQSAR system.[3,5] The claim is that the inclusion of local models in the hierarchy makes the overall accuracy of prediction greater, because the "systematic biases inherent in global model predictions are almost entirely removed" in the local models. Interestingly, Wood et al.[3] also suggest that including series models in the hierarchy does not help and sometimes hurts, perhaps because series contain too few compounds to extrapolate beyond the training set.

The hierarchical approach is complex and the source of the improvement of the hierarchical approach over global models is not necessarily clear, especially since not all the details of the approach are revealed in Wood et al.[3] In this paper, we directly compare the ability of global and local models to prospectively predict project compounds immediately after a model is build. We will do our study using a simple workflow and our own favored QSAR methods and descriptors. We aim to answer the following questions:

1. Is the local model best at predicting recent compounds also best at predicting new compounds?
2. Is the best local model systematically better than a global model?
3. Is a "consensus"[10,11] model a viable alternative to choosing the best local model?
4. Is there evidence that global models are systematically biased relative to local models?

METHODS

Training and Test Sets for Prospective Prediction. This discussion refers to Figure 1. Consider a set of molecules from project *P* tested on an off-target assay *X*. The molecules are ordered in terms of dates of testing on *X*. There will be a "date of model-building" for the global model (which varies from assay to assay). All molecules tested up to that date for *X* will be included in the global model. Let set *C* be the first *N* compounds in *P* tested for *X* after the date of model-building. Let set *B* be the last *N* compounds tested before that date. *A* is the set of all *P* compounds tested before *B*. The idea is to find out how to best prospectively predict *C*, given that we have a global model and a variety of local models. The idea that one can select a superior local model for predicting *C* assumes the following:

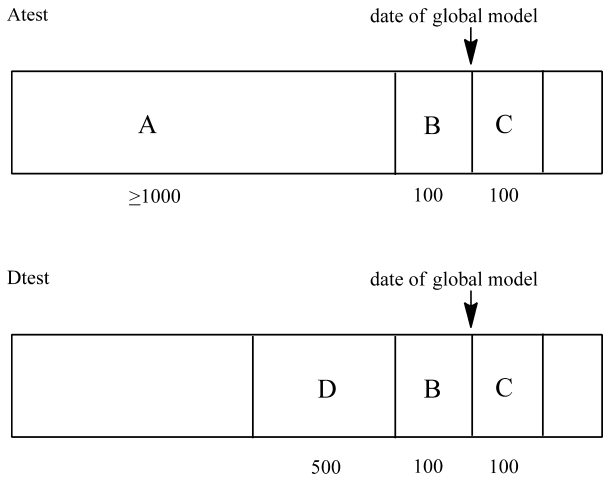


Figure 1. Bar representing compounds in a particular project *P* tested on a particular off-target assay *X* ordered by date of testing going from left to right with increasing time. The question is how to best predict *C*, compounds for *P* assayed after the last model is built: using a global model for *X* vs a variety of local models for *X*/*P*. There are two possible types of local (i.e., *P*-specific) models one could build: one using all compounds in project *P* (*A*test) or one using only recently tested compounds in project *P* (*D*test).

1. The best method/descriptor combination to predict *B* from *A* is also the best combination to predict *C* from *AB*.
2. The prediction of *C* by *AB* using the best method/descriptor combination of *B* predicted from *A* should be better than the prediction of *C* by the global model.

Here we investigated $N = 100$ for *B* and *C*. We would say 100 is a realistic number of compounds to predict in a project at one time, although it might be somewhat small for gathering stable statistics. (We also tried $N = 500$.) An alternative to taking a constant number of compounds for *B* and *C* would have been to make sets *B* and *C* be defined by time as was done in the work of Wood et al.[3] (for example, within 1 month before or after the global model was built). However, in some *X*/*P* combinations this would result in very few molecules and, in other combinations, very many; we felt using a constant number of compounds would make the tests more uniform among the *X*/*P* combinations.

Above we considered making project models using all the molecules associated with a project before a certain time, as in set *A*, which we will call the "*A*test experiment". We also consider building models that are "time local" as well as "project local". Here we consider a training set consisting of recently tested molecules (here the 500 project compounds that precede *B*) called set *D*. The discussion about the predictions to be made is the same as above, except that one substitutes "*D*" for "*A*" and "*DB*" for "*AB*". This is the "*D*test experiment".

Assays and Projects. Since we are interested in testing prospective prediction, this requires knowing dates of testing, and usually that requires using in-house data, which is proprietary. For this study *X* is among the off-target assays in Table 1. *P* can be one of 11 in-house projects, called Project01 to Project11.

In theory we could examine all combinations of *X* and *P*, but we eliminated combinations where set *A* would contain <1000 molecules or *D* < 500 molecules. That is, we are assuming that

dx.doi.org/10.1021/ci500084w | J. Chem. Inf. Model. 2014, 54, 1083−1092

Table 1. Data Sets for Global Models

activity	description	descriptor	QSAR method	number of compounds in $A + B$	mean ± stdev activity
HERG	binding to HERG channel $-\log(IC_{50})$ M	AP, DP	RF	195960	5.2 ± 0.76
CAV12	binding to calcium voltage-gated channel 1.2 $-\log(IC_{50})$ M	AP, DP	RF	54399	4.9 ± 0.43
NAV15	binding to sodium voltage-gated channel 1.5 $-\log(IC_{50})$ M	AP, DP	RF	49882	4.8 ± 0.40
2C9	inhibition of CYP 2C9 $-\log(IC_{50})$ M	AP, DP	RF	112289	4.8 ± 0.62
2D6	inhibition CYP 2D6 $-\log(IC_{50})$ M	AP, DP	RF	112250	4.5 ± 0.46
3A4	inhibition of CYP 3A4 $-\log(IC_{50})$ M	AP, DP	RF	112286	4.7 ± 0.67
TDI	time dependent inhibition for 3A4, log of ratio of IC_{50} with and without NADPH	AP, DP	RF	19827	0.61 ± 0.50
PXR	induction of CYP 3A4 of compound relative to induction by rifampicin expressed as percent	AP, DP, MOE_2D	RF	104498	42.4 ± 41.1
HPLC_LOGD	LOGD measured by HPLC	AP, DP	SVM	187931	2.7 ± 1.2

a reasonably large amount of data exists from which to make a model. Even with the elimination of many X/P combinations, we are examining two to three times more combinations than in Wood et al.[3]

QSAR Methods and Descriptors. To generate sets of project models for each assay/project combination, we used the following three QSAR methods:

1. Random forest (RF):[12] we are using a parallelized version of the original Brieman code.[13] Models used 100 trees. We routinely use prediction rescaling[14] to counteract the tendency of random forest predictions to compress the range of activity.
2. Linear kernel SVM as implemented in liblinear.[15]
3. Partial-least-squares as implemented in the R module *pls*.[16]

All models were run as regressions. Many data sets for the global model contain "qualified data" due to limits in the experiment, e.g. $IC_{50} > 30$ μM. These were treated as fixed numbers.

Whereas Wood et al.[3] used a common set of descriptors for all QSAR models, we are using the following four descriptor combinations:

1. AP, DP. This is the union of the original Carhart atom pairs,[17] and a donor–acceptor pair (called BP in the work of Kearsley et al.[18]).
2. TT, DT. This is the union of the original topological torsion[19] and the donor–acceptor torsion (called BT in the work of Kearsley et al.[18])
3. ECFP4. This is the circular fingerprint described by Rogers and Hahn.[20]
4. MOE_2D descriptors from the MOE package.[21]

Thus, there are 12 method/descriptor combinations for the local models.

Metric for Goodness of Prediction. Goodness of prediction is traditionally measured by R^2 (the higher the better) or RMSE (root-mean-square-error: the lower the better). Since the number of compounds being predicted is fairly small and may not cover the whole range of activity, R^2, which measures correlation only, may be misleading. We prefer to use RMSE, which looks at the numerical match of predicted and observed value. Since different Xs have different ranges in activity, to put all Xs on a single plot, one should normalize RMSE by dividing by the standard deviation of the observed activity for X in the data set from which the global model is calculated. We will call this RMSENORM. An RMSENORM of zero implies a perfect prediction. An RMSENORM of 1 indicates a prediction no better than guessing. It is possible to have RMSENORM > 1, i.e., the prediction is worse than guessing.

Global Models. Global models for each X were generated using all compounds tested on X before the cutoff date. These include all the project compounds included in the local models but also compounds from many other projects not in our project list. Information about these models is in Table 1. A good method/descriptor combination for building each global model was determined by cross-validation using only compounds tested before the date of model-building. In most cases the best method was RF. However, in one case (HPLC_LOGD), the best method was clearly SVM. Generally global models use the AP, DP combination. One (PXR) uses AP, DP, MOE_2D where the addition of MOE_2D descriptors improves the cross-validated prediction. Once calculated for each X, the global models were used without further modification.

Experimental Protocol. Here is the protocol for each X/P combination in the *A*test experiments:

1. Generate 12 models for A using all method/descriptor combinations. Use each model to predict B.
2. Generate 12 models for AB. Use each model to predict C.
3. Pick out the best model for B from A based on the lowest RMSE.
4. Pick out the best model for C from AB based on the lowest RMSE. This prediction is called BEST_LOCAL_C.
5. Record the RMSENORM for C from AB using the best method/descriptor combination of B from A in step 3. This prediction is called LOCAL_C_BY_BEST_LOCAL_B.
6. Generate the consensus prediction of C by averaging the predictions from the 12 models of AB in step 2. This prediction is called CONSENSUS_LOCAL_C.
7. Predict C with the global model of X. This prediction is called GLOBAL_C.

The protocol for the *D*test experiments is the same, except D is substituted for A and DB is substituted for AB.

RESULTS

There are 63 X/P combinations for the *A*test experiments and 72 X/P combinations for the *D*test experiments. All methods and all descriptors appear at least once in the best combinations. That implies that there is no one universally "good" or "bad" method or descriptor for local models, although on the average some method/descriptor combinations

dx.doi.org/10.1021/ci500084w | J. Chem. Inf. Model. 2014, 54, 1083–1092

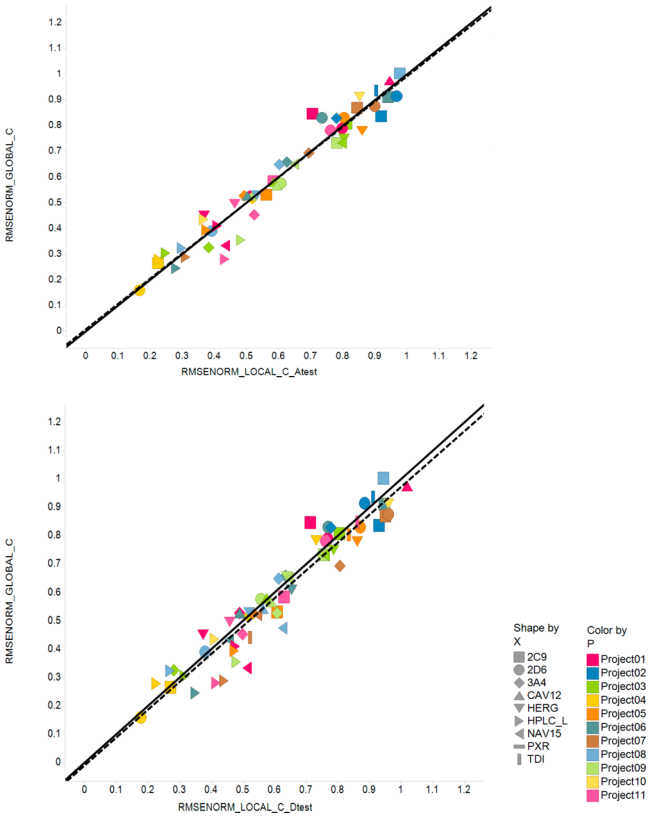


Figure 2. RMSENORM for prediction of *C* by the global model vs the prediction of *C* by the local model using the same method/descriptor combination. This excludes PXR because the global model contains an extra descriptor MOE_2D. (top) *A*test. (bottom) *D*test. In this and subsequent plots, the assays *X* are represented by shape, and the projects *P*, by color. The dashed line is the best linear fit, and the solid line is the diagonal.

are better than others. The method/descriptor combination (for *C* from *AB* or *DB*) that shows up as best most often is RF/APDP (21% of the time in the *A*test and 16% of the time in the *D*test). This is not a surprise since we usually have had good results with RF and APDP in the past. RF/MOE_2D is the second most frequent combination (16% in *A*test and 15% in *D*test).

How often does the best combination for *B* from *A* match the best combination of *C* from *AB* (*A*test) or *B* from *D* match *C* from *DB* (*D*test)? It is surprisingly infrequently: 23% of the time for *A*test and 18% for *D*test. This is clearly better than chance level of matching (~8%), but we would expect to do better than chance because, as noted in the above paragraph, some method/descriptor combinations are better than others. We can say that the best method/descriptor combination for the previous set of 100 compounds is also best for the next 100 compounds only a small minority of the time. Thus, the first assumption of the "best local model" idea seems dubious.

Figures 2−10 will show the RMSENORM of one method of prediction vs another. Each point represents an *X*/*P* combination. *X*s will be distinguished by shape and projects by color. One important observation is that, in all plots, all *X*/*P* combinations fall more or less on the same line, meaning that the relationship between global and local models is not strongly dependent on *X* or *P*. It makes sense to find the best least-squares fit through all the points, and in the plots this is shown as a dashed line. If the methods being compared in the plot were equivalent, we would see the points equally spread above and below the diagonal (solid black line) and the dashed line close to the diagonal. If one method was systematically better than another, we would see the points above or below the diagonal and the least-squares line shifted accordingly; lower RMSENORM is "better prediction."

One can quantitate the deviation from the diagonal for each *X*/*P* combination by Diff = RMSENORM(Prediction1) − RMSENORM(Prediction2).

To summarize the deviation for all *X*/*P* combinations in a plot, one calculates the mean of Diff (called meanDiff) and one can calculated the scatter for all *X*/*P* combinations by taking the standard deviation of Diff (called stdevDiff). A meanDiff <0 implies that Prediction1 is better. MeanDiff and stdevDiff are listed in Table 2. It should be noted that the meanDiff in Table 2 is often small compared to stdevDiff. That is, systematic shifts

Table 2. MeanDiff for Pairs of Models in the Figures

figure ref	experiment	pred1	pred2	meanDiff ± stdevDiff pred1 minus pred2
Figure 2 top	*A*test	GLOBAL_C	LOCAL_C same method/descriptor	−0.001 ± 0.055
Figure 2 bottom	*D*test	GLOBAL_C	LOCAL_C same method/descriptor	−0.021 ± 0.062
Figure 3 top	*A*test	GLOBAL_C	LOCAL_C_BY_BEST_LOCAL_B	0.004 ± 0.075
Figure 3 bottom	*D*test	GLOBAL_C	LOCAL_C_BY_BEST_LOCAL_B	−0.036 ± 0.114
Figure 4 top	*A*test	GLOBAL_C	BEST_LOCAL_C	0.058 ± 0.062
Figure 4 bottom	*D*test	GLOBAL_C	BEST_LOCAL_C	0.036 ± 0.070
Figure 5 top	*A*test	GLOBAL_C	CONSENSUS_LOCAL_C	0.031 ± 0.074
Figure 5 bottom	*D*test	GLOBAL_C	CONSENSUS_LOCAL_C	0.001 ± 0.094
Figure 6 top	*A*test	CONSENSUS_LOCAL_C	BEST_LOCAL_C	0.027 ± 0.048
Figure 6 bottom	*D*test	CONSENSUS_LOCAL_C	BEST_LOCAL_C	0.036 ± 0.064
Figure 7 top	*A*test	CONSENSUS_LOCAL_C	LOCAL_C_BY_BEST_LOCAL_B	−0.027 ± 0.055
Figure 7 bottom	*D*test	CONSENSUS_LOCAL_C	LOCAL_C_BY_BEST_LOCAL_B	−0.036 ± 0.089
Figure 8 top	*A*test	LOCAL_C_BY_BEST_LOCAL_B	LOCAL_C_BY_BEST_LOCAL_B	0.035 ± 0.091
Figure 8 bottom	*D*test	CONSENSUS_LOCAL_C	CONSENSUS_LOCAL_C	0.022 ± 0.041
Figure 9 top	*S*test	GLOBAL_C	LOCAL_C same method/descriptor	−0.058 ± 0.073
Figure 9 middle	*S*test	GLOBAL_C	LOCAL_C_BY_BEST_LOCAL_B	−0.032 ± 0.078
Figure 9 bottom	*S*test	GLOBAL_C	CONSENSUS_LOCAL_C	−0.001 ± 0.059
Figure 10	*A*test RF/MOE_2D	GLOBAL_C RF/MOE_2D	LOCAL_C RF/MOE_2D	0.040 ± 0.097

dx.doi.org/10.1021/ci500084w | *J. Chem. Inf. Model.* 2014, 54, 1083−1092

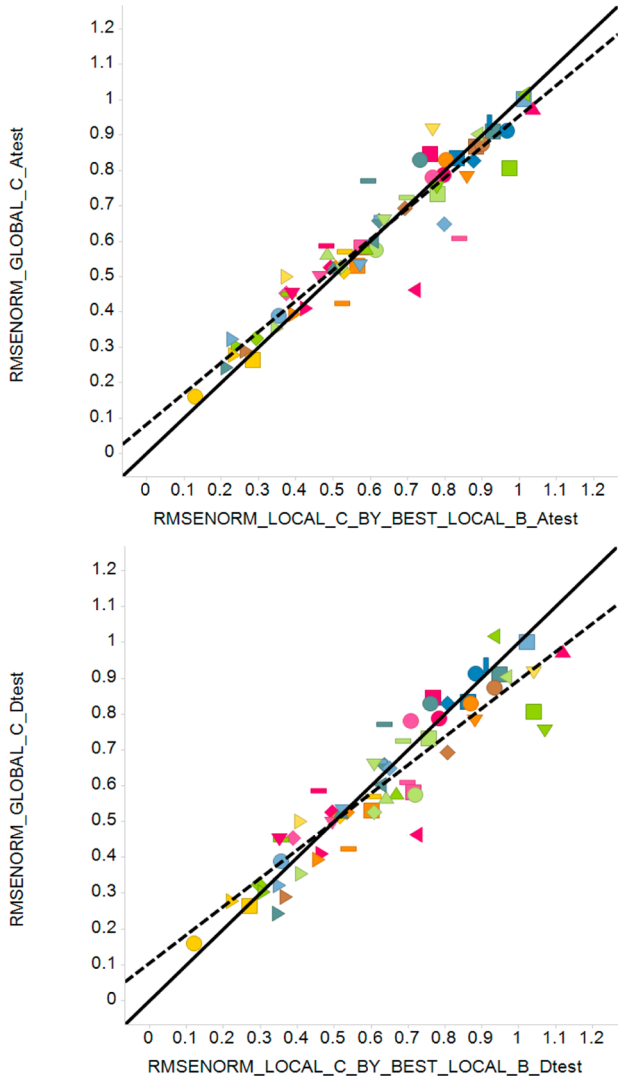


Figure 3. (top) RMSENORM for prediction of *C* by the global model vs the prediction of *C* from *AB* using the best method/descriptor combination for the prediction of *B* from *A*. (bottom) Prediction of *C* by the global model vs the best prediction of *C* from *DB* using the best method/descriptor combination for the prediction of *B* from *D*.

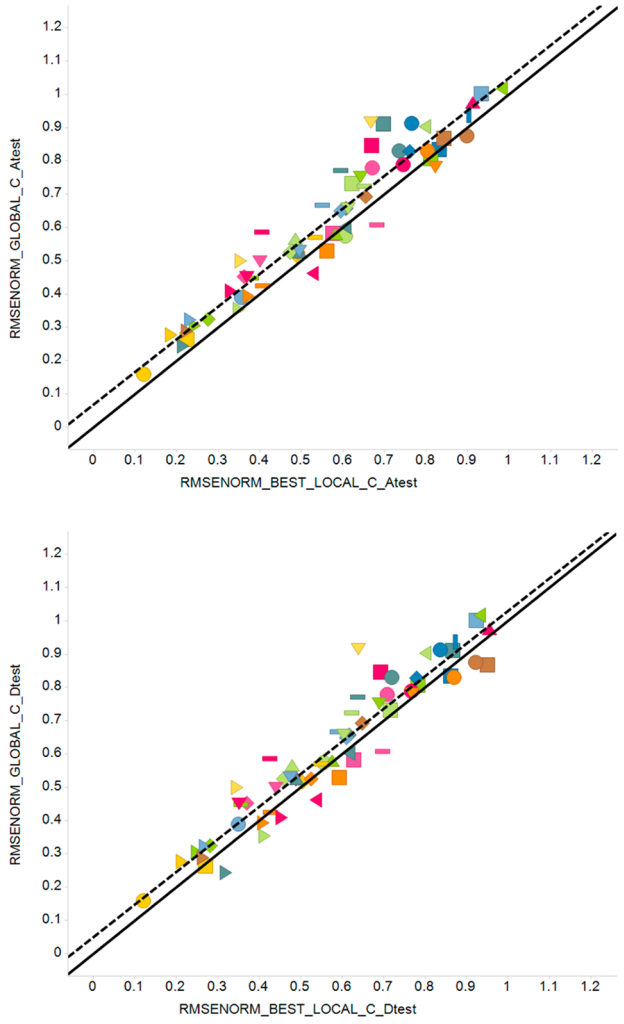


Figure 4. (top) RMSENORM for prediction of *C* by the global model vs the prediction of *C* from *AB* using the best method/descriptor combination of *C* from *AB*. (bottom) Prediction of *C* by the global model vs the prediction of *C* from *DB* using the best method/descriptor combination of *C* from *DB*. Note that this involves cheating since picking the best involves knowing the activities of *C* in advance.

between methods of prediction can be small compared to the variation among individual *X*/*P* combinations.

Figure 2 shows the baseline comparison. How much does the prediction of C by the global model differ from the prediction of C by the local model using the same method/descriptor combination? Figure 2 (top) shows this for *A*test, and Figure 2 (bottom) shows this for *D*test. In Table 2 meanDiff shows there is effectively no difference between global and local models for *A*test and a very small shift toward better prediction by the global model for *D*test.

What about the idea of getting better predictions by selecting local models? Figure 3 compares the global and the selected local model. That is, it is comparing step 7 of the protocol vs step 5. Figure 3 (top) shows the RMSENORM of the prediction of *C* by the global model vs RMSENORM of the prediction of *C* from by the local *AB* model using the best method/descriptor combination of *B* from *A*. Figure 3 (bottom) shows the equivalent for *C* from the local *DB*. The meanDiff for the top figure is very small and the meanDiff for the bottom figure is slightly negative. That is, there does not seem to be a systematic advantage of using a selected local model vs the global model, contrary to the expectations.

In Figure 4, we compare step 7 from the protocol to step 4. In Figure 4 (top) one can show a clear systematic decrease in RMSENORM in the best local model of *C* from *AB* relative to the prediction of *C* from the global model. Figure 4 (bottom) shows the equivalent for *C* from *DB*. We show this as an example of the maximum possible predictivity from a local model. It is important to note, however, that this is involves cheating, since to know which local model was best for *C*, we would have to know the activities of *C* in advance.

If the suggestion to select the best model using recent data fails to produce a significant improvement, will using a consensus model be better? In Figure 5, we are comparing the global model to a consensus model. That is, we are comparing step 7 in the protocol to step 6. Figure 5 (top) RMSENORM of the prediction of *C* by the global model vs the prediction of *C* by the consensus of local models of *AB*. Figure 5 (bottom) shows the equivalent for the consensus models of *DB*. Note that there is no cheating involved in predicting *C* by *AB* or *DB* through consensus since we are doing no selecting,

dx.doi.org/10.1021/ci500084w | *J. Chem. Inf. Model.* 2014, 54, 1083−1092

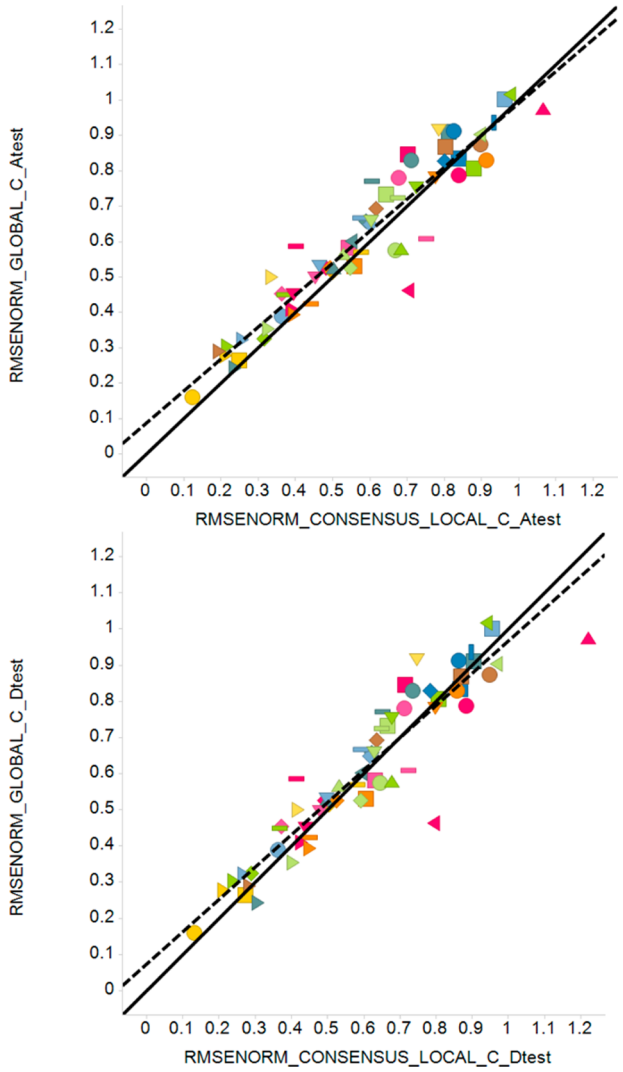


Figure 5. (top) RMSENORM for the prediction of *C* by the global model vs the consensus prediction of *C*, where the "consensus" is the average prediction of 12 local models of *AB*. (bottom) Equivalent for the consensus models of *DB*.

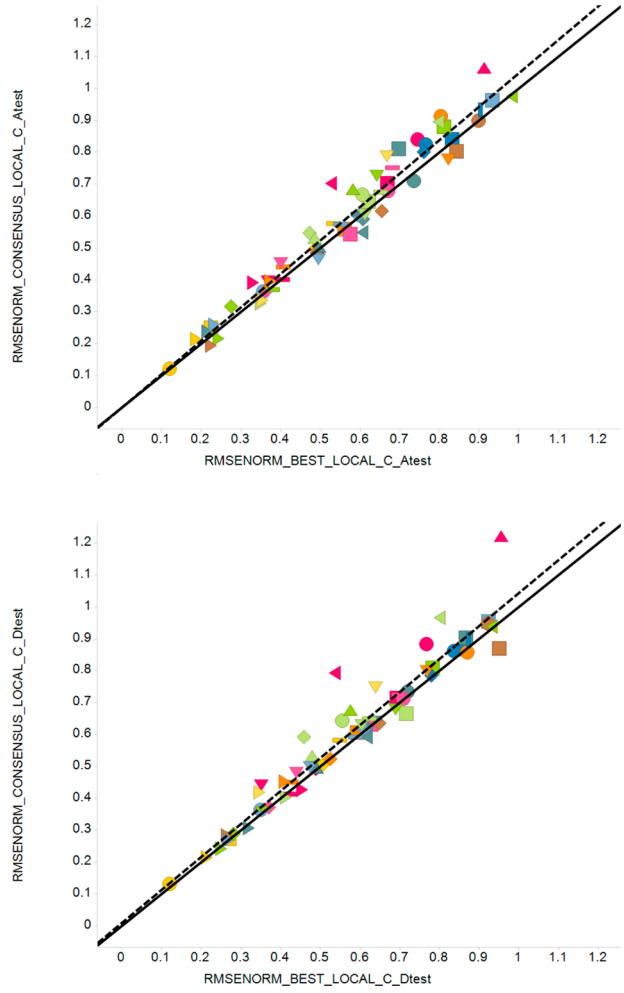


Figure 6. (top) Comparison of the consensus model for local *C* from *AB* vs the best local model of *C* from *AB*. (bottom) Equivalent of for local *C* from *DB*.

just averaging over the 12 models. There is only a very small shift toward better prediction by the consensus model vs the global model indicated as indicated by meanDiff.

Figure 6 compares the consensus model to the best local model. This is comparing step 5 of the protocol against step 6. Figure 6 (top) shows that the consensus model for *C* from *AB* vs the best model for *C* from *AB*. Figure 6 (bottom) shows the equivalent for *C* from *DB*. The meanDiff indicates that the best local model for *C* may be slightly better than the consensus. This is consistent with previous observations that consensus models in QSAR are not significantly better than the best single model[10,11] and may be worse.

Figure 7 shows the direct comparison of the prediction of consensus model vs the best local model using recent data. This is comparing step 5 from the protocol against step 6. Figure 7 (top) shows the equivalent for *A*test, and Figure 7 (bottom) shows the equivalent for *D*test. As might be expected from comparing Figures 2 and 4, the consensus model is slightly better. While this suggests that consensus is a better use of local models than selecting the best local model based on the prediction of recent molecules, it does not necessarily reflect on the question of whether local models are better than global models since there are no consensus global models to directly compare against.

Figure 8 compares *A*test to *D*test. Figure 8 (top) compares the prediction of local *C* from *AB* using the best method/descriptor combination of *B* from *A* vs local *C* from *DB* using the best method/descriptor combination of *B* from *D*. This is step 5 from the protocol. Figure 8 (bottom) compares the prediction of *C* from the consensus local model from *AB* to the prediction of *C* from the consensus local model from *DB*. This is step 6 from the protocol. The meanDiff indicates that *AB* models are slightly better than *DB* models in predicting *C*, consistent with what we see in Figure 2. This might argue the better course is to cover more chemical space in a model than to be time local to the compounds being predicted.

An entire set of parallel experiments (both *A*test and *D*test) was done where the size of *B* and *C* was set to 500 instead of 100. This was done to get better statistics and so the range of predicted activity for *C* would be closer to the range of observed activities in the entire assay. The results of these experiments are qualitatively very similar to what we have presented here. The only difference is that there are fewer *X*/*P* combinations (because fewer combinations have enough compounds in *A* and/or *D*) and that the scatter in the plots is less.

dx.doi.org/10.1021/ci500084w | J. Chem. Inf. Model. 2014, 54, 1083−1092

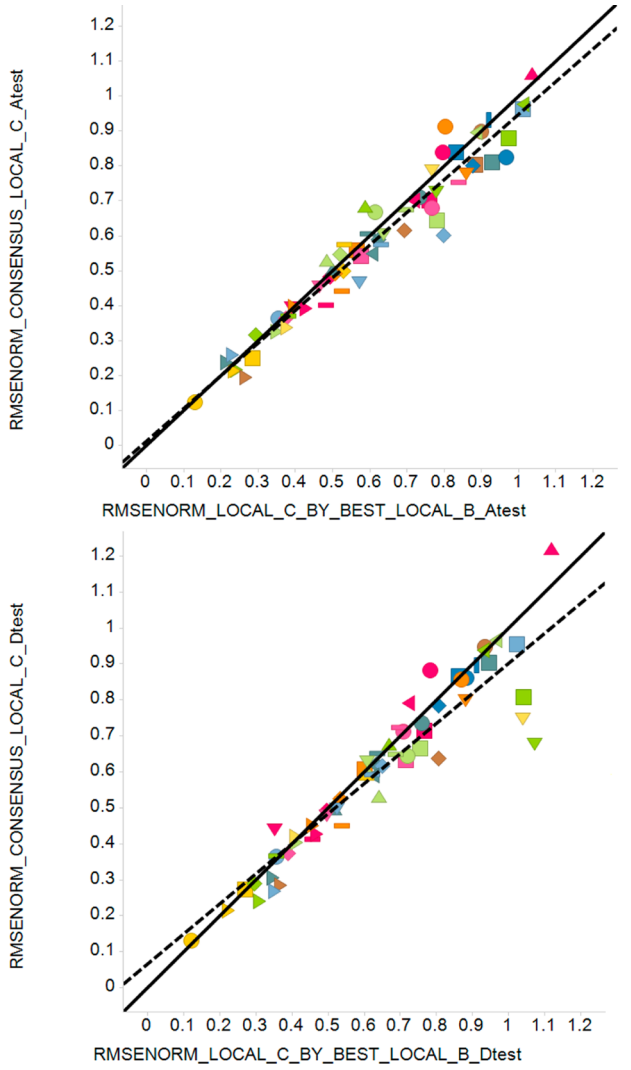


Figure 7. (top) RMSENORM for the consensus local model for *C* vs the prediction of *C* from *AB* using the best method/descriptor combination for predicting *B* from *A*. (bottom) Equivalent for *C* from *DB* using the best method/descriptor combination for predicting *B* from *D*.

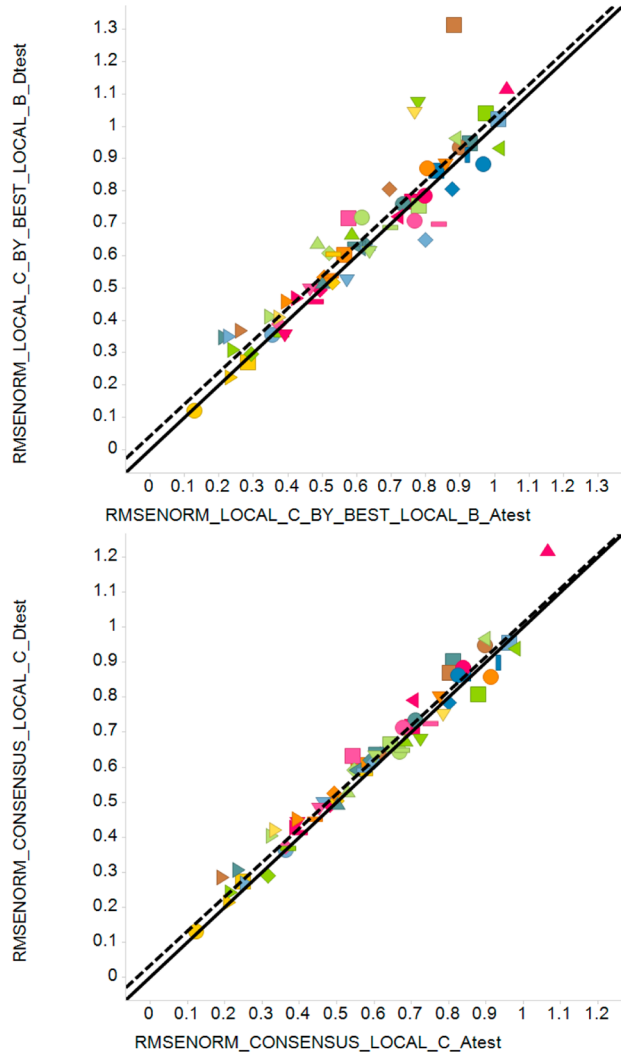


Figure 8. (top) Comparison of the prediction of local *C* from *AB* using the best method/descriptor combination for *B* from *A* vs the prediction of *C* from *DB* using the best method/descriptor combination of *B* from *D*. (bottom) Comparison of the prediction of *C* by the consensus local model of *AB* vs the prediction of *C* by the local consensus model of *DB*.

A believer in the superiority of local models could criticize the work presented above on the basis that our local models are not local enough. Both the local and global models contain more than one chemical series, so both types of models could be equally biased. If we had made local models containing a single series, the thinking goes, the prediction of local models on compounds in the same series would have been better. To address that idea we made a modification of *A*test, called *S*test, such that only a single series was considered per project. We clustered the compounds for each project using the Butina algorithm,[22] the ECFP4 descriptor, and a Tanimoto similarity cutoff of 0.5. The largest cluster of compounds for the project was taken as the series of interest for that project. There were 16 *X*/*P* combinations where *A* has ≥150 compounds and *B* and *C* have 100 compounds, all from that series. Selected plots for *S*test are shown in Figure 9 (top), (middle), and (bottom), corresponding to Figures 2 (top), 3 (top), and 5 (top), respectively. Clearly, from these plots there is no superiority of prediction of the series model on series compounds compared to the prediction of the global model on series compounds. Since there are so few points one cannot be sure, but the opposite seems to be the case. If so, this would be consistent with the observation of Wood et al.[3] that series models may give slightly poorer predictions than local models.

Two additional experiments were done to test possible circumstances where local RF models would appear better than global RF models. The reason for these experiments will be made clear in the Discussion. The first experiment is not to use prediction rescaling in RF for either global or local models in *A*test. In practice this had no discernible effect on the relative goodness of global and local models. Plots from that experiment are not distinguishable from Figures 2−5 (top).

The second experiment had to do with the type of chemical descriptors (property vs substructure) used in the QSAR models. In this experiment all global models were rebuilt using RF/MOE_2D. The RMSENORM for the prediction of *C* from the RF/MOE_2D global models was compared with the RMSENORM of the prediction of *C* from the local *AB* model using RF/MOE_2D. That plot is shown in Figure 10. This is analogous to Figure 2 (top), where the global and local models

dx.doi.org/10.1021/ci500084w | J. Chem. Inf. Model. 2014, 54, 1083−1092

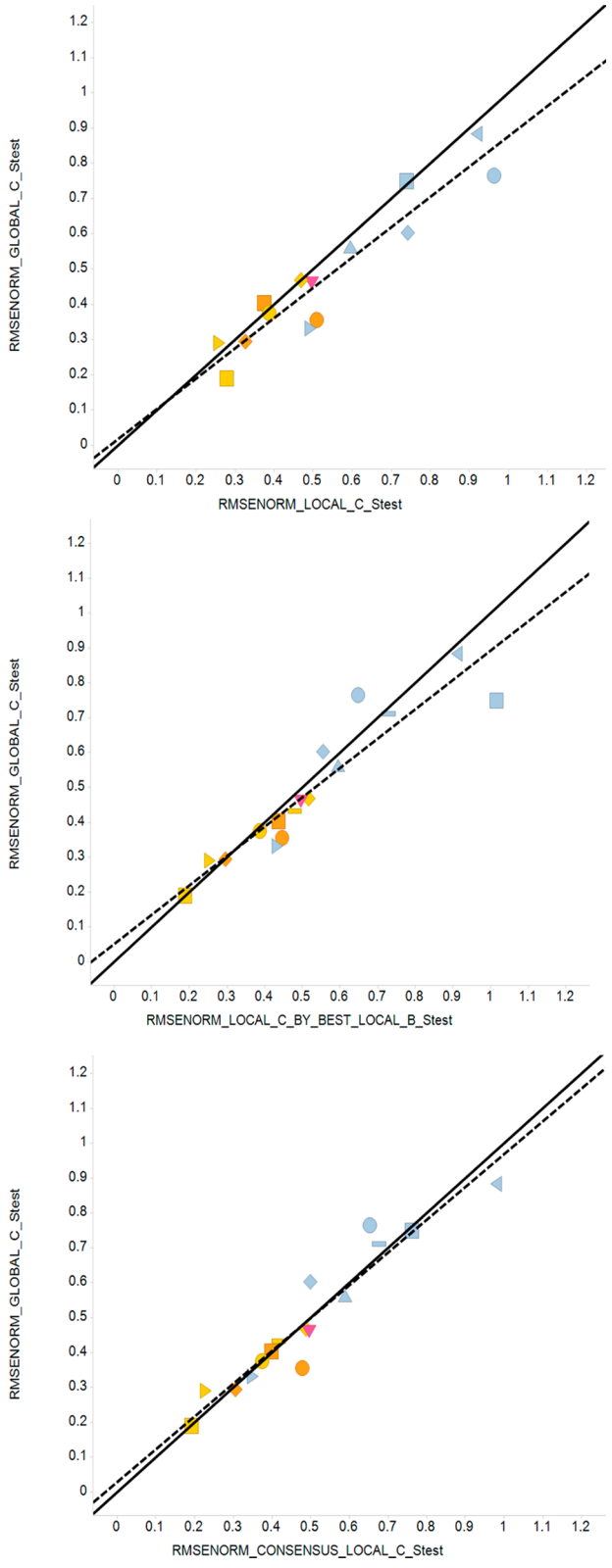


Figure 9. Prediction of C by the global model for compounds in a series vs prediction by local models made from the same series. (top) Global models and series models using the same method/descriptor combination. (middle) Series model predicts *C* from *AB* using the best method/descriptor combination for the prediction of *B* from *A*. (bottom) Series models using consensus prediction of *C*, where the consensus is the average prediction of 12 series models of *AB*.

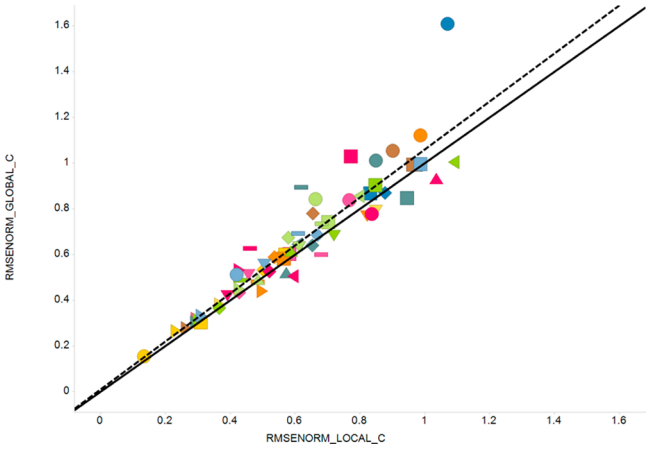


Figure 10. RMSENORM for prediction of *C* by the global model vs the prediction of *C* by the local model, all models using RF and MOE_2D.

are using the same method/descriptor combination, except here the combination is always RF/MOE_2D. In this plot there is a discernible advantage to the local models, with meanDiff similar in magnitude to that seen in Figures 4 (top) and 5 (top). Removing the apparent outlier (2D6/Project02) does not change the conclusion.

DISCUSSION

In our exercise we are comparing various ways of prospectively predicting project compounds using global vs local (project) models. In all our plots comparing methods, individual points (*X*/*P* combinations) are scattered above or below the diagonal, so for any specific *X*/*P* combination, a local model could be slightly better than a global model, or the opposite could be true. However, the overall trend as displayed by the least-squares fit line through the points is never very far from the diagonal. Another way of expressing the same result is that the absolute value of meanDiff is never large compared to stdevDiff, that is the overall difference among two methods is not large compared to the scatter in the individual examples. The largest meanDiff is seen in Figure 4 (top), which involves cheating by knowing the prospective activities in advance, and even then, the absolute meanDiff is smaller than stdevDiff. Thus, all our global models are roughly as good as local models for prospective prediction. It does not seem to matter significantly if the local models are time local in addition to being project local (i.e., *D*test vs *A*test) or even of the same series. Trying multiple method/descriptor combinations, either to pick the best local model based on recently tested compounds, or to generate a consensus model, also does not seem to make an appreciable difference. Therefore, using a single well-updated global model per off-target assay *X* seems sufficient. There is added complexity in generating a set of project- or series-specific models for *X*, and complexity in having to decide which model should be applied to predict the *X* activity of a new molecule. Given our results, that complexity seems hard to justify.

Why is the best project model for *B* from *A* (or from *D*) not indicative of the best model of *C* from *AB* (or from *DB*), i.e. why is the prediction of recent data not indicative of the prediction of future data? One general type of explanation is that the predictions are much more sensitive to which training set one is using (even if the difference is as small as 100

dx.doi.org/10.1021/ci500084w | J. Chem. Inf. Model. 2014, 54, 1083−1092

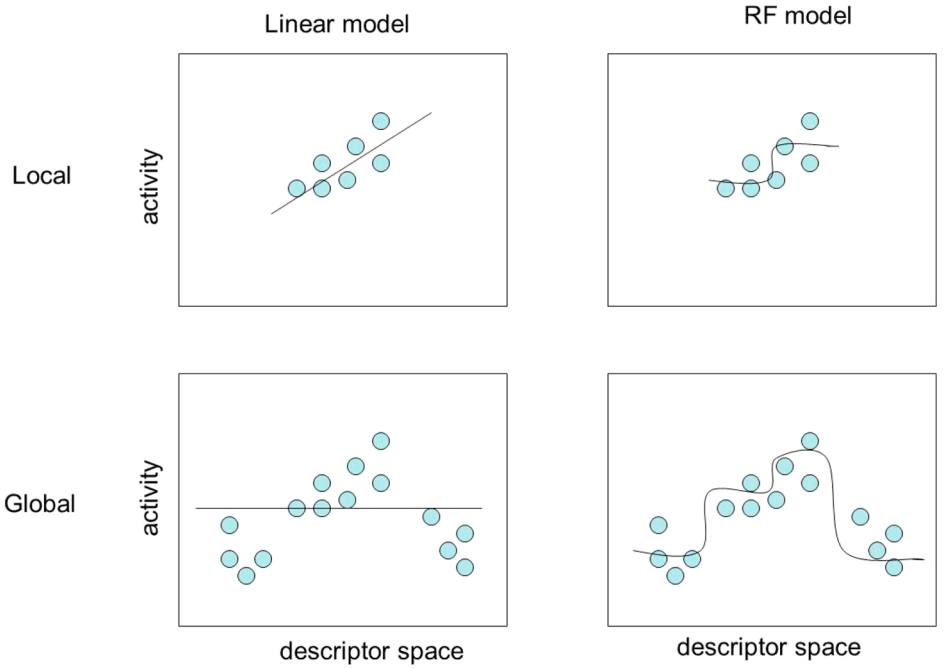


Figure 11. Cartoon illustrating expectations in building local vs global models. Methods like random forest include the local model as part of the global model.

molecules out of a few thousand) than which method/descriptor combination one uses. If that is true, one general consequence is that it might be futile to find the absolute best method/descriptor combination for any kind of QSAR model based on recent data and expect that model to be the best for the future. Picking one "good enough" method/descriptor combination is probably the best that can be done in practice.

On the basis of studies not discussed in this paper, we do agree with Wood et al.[3] about the importance of keeping global models updated. We cannot say we have refuted the *observation* of Wood et al.[3] that a hierarchical approach using global, local, and series models can slightly outpredict an updated global for their off-target assays and projects, since we did not follow the hierarchical approach. However, we can say our results do not support the *assumption* that the local model that best predicts recent data will best predict future data. Also our results do not support the *explanation* offered by Wood et al.[3] that local models are superior to global models because local models lack the "systematic bias" inherent in global models. We interpret this to mean that, while global models might predict trends in activity correctly, they would not get the numerical values correct and we would expect RMSENORM to be systematically higher for global models. However, we detect no such bias using our data sets, QSAR methods, and descriptors.

Why do Wood et al.[3] detect bias in global models while we do not? Since their efforts and ours differ in many respects (data sets, descriptors, workflow, etc.), it may not be possible to definitely answer that question. However, we can manipulate our own methods and descriptors to try to find circumstances where local models look better than global models. We will examine two possibilities. One is our use of prediction rescaling in RF, which is not likely to have been used by Wood et al.,[3] since it is not a common practice. Perhaps not rescaling would systematically raise the RMSENORM of the global models, but not raise the RMSENORM of the local models as much (if they lacked bias). The RMSENORM of the non-RF models, of course, would not change. On the average, then, the local models might start to look better. As was mentioned in the RESULTS, however, removing prediction rescaling had a negligible effect, so this is a less plausible explanation.

Another possible explanation has to do with the type of descriptors used for the models. The descriptor set used by Wood et al.[3] is dominated by property descriptors (where there is one number per molecule, e.g. LOGP, number of hydrogen bond donors, molecular weight, etc.). In contrast, our descriptor set is dominated by substructure (sometimes called fingerprint) descriptors, i.e. the count of the occurrences of a variety of substructures in a given molecule. When we made both global and local models with MOE_2D descriptors (property descriptors), local models looked slightly better than the global models. Thus, the descriptor type could be a plausible explanation to explore further.

Many chemists have a strong expectation that QSAR models trained on compounds more local to the compounds being predicted will be better than models trained on large diverse data sets. However, in practice we can show that the expectation is false; local and global models are roughly equivalent. Why, then, is the expectation so strong? It may have to do with subconsciously thinking that all "models" will behave like "linear models." This is illustrated in Figure 11. Imagine that a series of compounds from project *P* forms a clear linear activity trend in chemical space (upper left). The model is indicated by a line through the data. Adding "irrelevant" compounds to the model outside the series (lower left) will clearly spoil the linear model for predicting the original series. In contrast, if one uses a piecewise method like random forest (upper right), the local model will fit the series. Adding more compounds outside the series has no deleterious effect because the global model (lower right) includes the local model. Thus, the predictions of the local and global models for the project will be equivalent. This may explain why RF seems to give a better match of local vs global predictions than SVM or PLS as noted by Helgee et al.[2] Another possible explanation has to do with the fact that SVM and PLS require normalization of

dx.doi.org/10.1021/ci500084w | J. Chem. Inf. Model. 2014, 54, 1083−1092

descriptors (scaling descriptors in a range 0 to 1 or by the mean and standard deviation) where RF does not. We would expect the normalization to depend heavily on the composition of the training set, and this would especially be an issue with property descriptors. However, even Helgee et al. noticed that it is possible to get good global models using SVM and PLS. Our SVM global model of HPLC_LOGD is an example.

CONCLUSION

1. We tested the idea that one can select a local model based on which QSAR method/descriptor combination best predicted recent compounds, and that this model will be systematically more predictive than an updated global model. This does not seem to be true for the examples and method/descriptor combinations we tried. Method/descriptor combinations best for the recent past were not best for the immediate future, and we saw no advantage for selected local models.
2. We also tested the idea that a consensus local model built from the average prediction of many local models built with different method/descriptor combinations would be more predictive than a global model. There might be an advantage to the consensus model over the selected local model, but the advantage is small. Thus generating multiple local models using different method/descriptor combinations does not seem cost-effective.
3. In all cases, global models were effectively as good as local models in predicting project compounds, and it is hard to justify the complexity of generating project- or series-specific models.
4. The type of chemical descriptors used (e.g., property vs substructure) may affect the relative predictivity of global vs local models.

ASSOCIATED CONTENT

Supporting Information

Table of RMSENORM for *A*test, *D*test, and *S*test. 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

Joseph Shpungin wrote the parallelized version of random forest that we use here. A large number of Merck biologists, over many years, generated the data for examples used in the paper.

REFERENCES

(1) Feher, M.; Ewing, T. Global or local QSAR: Is there a way out? *QSAR Comb. Sci.* **2009**, *28*, 850−855.
(2) Helgee, E. A.; Carlsson, L.; Boyer, S.; Norinder, U. Evaluation of quantitative structure-activity relationship modeling strategies: local and global models. *J. Chem. Inf. Model.* **2010**, *50*, 677−689.
(3) 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.
(4) Buchwald, F.; Girschick, T.; Seeland, M.; Kramer, S. Using local models to improve (Q)SAR predictivity. *Mol. Inf.* **2011**, *30*, 205−218.
(5) Davis, A. M.; Wood, D. J. Quantitative structure-activity relationship models that stand the test of time. *Mol. Pharmaceutics* **2013**, *10*, 1183−1190.
(6) Gua, R.; Dutta, D.; Jurs, P. C.; Chen, T. Local lazy regression: making use of the neighborhood to improve QSAR predictions. *J. Chem. Inf. Model.* **2006**, *46*, 1836−1847.
(7) Zhang, S.; Golbraikh, A.; Oloff, S.; Kohn, H.; Tropsha, A. A novel automated lazy learning QSAR (ALL-QSAR) approach: method development, applications, and virtual screening of chemical databases using validated ALL-QSAR models. *J. Chem. Inf. Model.* **2006**, *46*, 1984−1995.
(8) Zhang, H.; Ando, H. Y.; Chen, L.; Lee, P. H. On-the-fly selection of a training set for aqueous solubility prediction. *Mol. Pharmaceutics* **2007**, *4*, 489−497.
(9) Sommer, S.; Kramer, S. Three data mining techniques to improve lazy structure-activity relationships for noncongeneric compounds. *J. Chem. Inf. Model.* **2007**, *47*, 2035−2043.
(10) Hewitt, M.; Cronin, M. T. D.; Madden, J. C.; Rowe, P. H.; Johnson, C.; Obi, A.; Enoch, S. J. Consensus QSAR models: do the benefits outweight the complexity. *J. Chem. Inf. Model.* **2007**, *47*, 1460−1468.
(11) Riniker, S.; Fechner, N.; Landrum, G. A. Heterogeneous classifier fusion for ligand-based virtual screening: or, how decision making by committee can be a good thing. *J. Chem. Inf. Model.* **2013**, *53*, 2829−2836.
(12) 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.
(13) Breiman, L. Random forests. *Machine Learning* **2001**, *45*, 5−32.
(14) Sheridan, R. P. Using random forest to model the domain applicability of another random forest model. *J. Chem. Inf. Model.* **2013**, *53*, 2837−2850.
(15) *LIBLINEAR—A Library for Large Linear Classification*. http://www.csie.ntu.edu.tw/~cjlin/liblinear/, last accessed March 22, 2014.
(16) *Package PLS*. http://cran.r-project.org/web/packages/pls/index.html, last accessed March 22, 2014.
(17) 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.
(18) 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. Inform. Comp. Sci.* **1996**, *36*, 118−27.
(19) Nilakantan, R.; Bauman, N.; Dixon, J. S.; Venkataraghavan, R. Topological torsions: a new molecular descriptor for SAR applications. comparison with other descriptors. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 82−85.
(20) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Mod.* **2010**, *50*, 742−754.
(21) *Molecular Operating Environment (MOE)*, Version 2008, release 10; Chemical Computing Group: Montreal, Canada, 2009. www.chemcomp.com, last accessed March 22, 2014.
(22) Butina, D. Unsupervised data base clustering based on Daylight's fingerprint and Tanimoto similarity: A fast and automated way to cluster small and large data sets. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 747−750.