JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

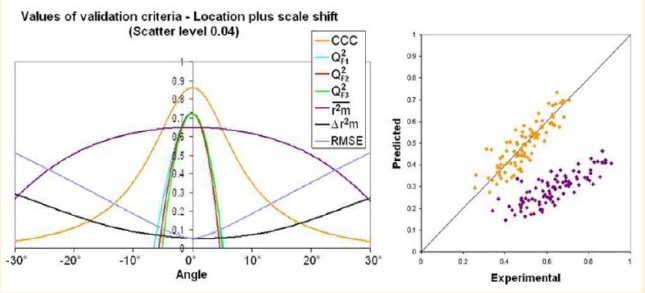
Real External Predictivity of QSAR Models. Part 2. New Intercomparable Thresholds for Different Validation Criteria and the Need for Scatter Plot Inspection

Nicola Chirico and Paola Gramatica*

QSAR Research Unit in Environmental Chemistry and Ecotoxicology, Department of Theoretical and Applied Sciences, University of Insubria, Via Dunant 3, 21100, Varese, Italy

Supporting Information

ABSTRACT: The evaluation of regression QSAR model performance, in fitting, robustness, and external prediction, is of pivotal importance. Over the past decade, different external validation parameters have been proposed: Q^2_{F1}, Q^2_{F2}, Q^2_{F3}, $\overline{r^2_m}$, and the Golbraikh–Tropsha method. Recently, the concordance correlation coefficient (CCC, Lin), which simply verifies how small the differences are between experimental data and external data set predictions, independently of their range, was proposed by our group as an external validation parameter for use in QSAR studies. In our preliminary work, we demonstrated with thousands of simulated models that CCC is in good agreement with the compared validation criteria (except r^2_m) using the cutoff values normally applied for the acceptance of QSAR models as externally predictive. In this new work, we have studied and compared the general trends of the various criteria relative to different possible biases (scale and location shifts) in external data distributions, using a wide range of different simulated scenarios. This study, further supported by visual inspection of experimental vs predicted data scatter plots, has highlighted problems related to some criteria. Indeed, if based on the cutoff suggested by the proponent, $\overline{r^2_m}$ could also accept not predictive models in two of the possible biases (location, location plus scale), while in the case of scale shift bias, it appears to be the most restrictive. Moreover, Q^2_{F1} and Q^2_{F2} showed some problems in one of the possible biases (scale shift). This analysis allowed us to also propose recalibrated, and intercomparable for the same data scatter, new thresholds for each criterion in defining a QSAR model as really externally predictive in a more precautionary approach. An analysis of the results revealed that the scatter plot of experimental vs predicted external data must always be evaluated to support the statistical criteria values: in some cases high statistical parameter values could hide models with unacceptable predictions.



INTRODUCTION

Quantitative Structure–Activity Relationships (QSAR) model validation is fundamental to ensure the reliability of data predicted by the models. Because the main purpose of QSAR is to quickly and accurately estimate a property of interest for an untested set of compounds, it is important to have a simple means of correctly setting user expectations of model performance for predictivity on new chemicals, never used in the model development (external validation).[1−3] In spite of the recent works on external validation,[1−7] the topic is still open, and the problem in QSAR modeling has not yet been completely solved, though many techniques have been proposed to validate models[1,8−16] with various preferences by the different users and varying degrees of success. In our recent preliminary paper,[16] we compared the performance of the more widely used external validation parameters[1,8−16] using thousands of simulated QSAR models that had previously been verified for their internal quality ($R^2 > 0.7$, $Q^2_{LOO} > 0.6$, $|R^2 - Q^2_{LOO}|<0.1$): a necessary condition for any QSAR model but not sufficient.[1−4] As an additional criterion for the external validation of QSAR models, we proposed the use of a conceptually simpler statistical parameter that basically verifies the agreement of experimental and predicted data: the concordance correlation coefficient (CCC) of Lin.[17] Some problems on the more discrepant criterion r^2_m,[13] verified on both axis dispositions (now r^2_m and r'^2_m[14]) were highlighted. In the previous comparison, we used for each criterion the threshold values defined by the respective proponent (0.5 for r^2_m metrics) or those more commonly applied in current practice by the most prudent QSAR modelers (0.6 for Q^2_{Fn}). For CCC, we proposed a reasonable cutoff value of 0.85, arbitrarily defined by verifying scatter plots on external data, which were acceptable from our QSAR experience. Using thousands of simulated models (thus, not just on a single data set) and also on real examples, CCC was found to be in good agreement with the other studied validation measures (about 96%, except r^2_m metrics). When found in disagreement, it showed to be the most precautionary criterion (with the proposed cutoff)

Received: February 10, 2012

   dx.doi.org/10.1021/ci300084j | J. Chem. Inf. Model. XXXX, XXX, XXX–XXX

Table 1. Most Widely Used Formulas for External Validation of QSAR Models

1) $R^2_{EXT} = 1 - \frac{\sum_{i=1}^{n_{EXT}} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n_{EXT}} (y_i - \bar{y})^2}$ Determination coefficient of the prediction set external data.
Important note: in this formula, $\hat{y}_i$ is the predicted value calculated using the regression of the predicted and experimental data of the prediction set, while in the subsequent formulas the $\hat{y}_i$ value is calculated using the QSAR model.
2) $Q^2_{F1} = 1 - \frac{\sum_{i=1}^{n_{EXT}} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n_{EXT}} (y_i - \bar{y}_{TR})^2}$ (ref. 9)
3) $Q^2_{F2} = 1 - \frac{\sum_{i=1}^{n_{EXT}} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n_{EXT}} (y_i - \bar{y}_{EXT})^2}$ (ref. 10)
4) $Q^2_{F3} = 1 - \frac{\left[\sum_{i=1}^{n_{EXT}} (y_i - \hat{y}_i)^2\right] / n_{EXT}}{\left[\sum_{i=1}^{n_{TR}} (y_i - \bar{y}_{TR})^2\right] / n_{TR}}$ (ref. 11 and 12)
5) $r^2_m = r^2\left(1 - \sqrt{r^2 - r_0^2}\right)$ $\overline{r^2_m} = \frac{r^2_m + r'^2_m}{2}$ $\Delta r^2_m = \left\lvert r^2_m - r'^2_m \right\rvert$ (ref. 14 and 15)
r^2 and r_0^2 are respectively the determination coefficients of the regression function calculated using the experimental and the predicted data of the prediction set, forcing respectively the origin of the axis (r_0^2) or not (r^2).
r^2_m is calculated using the experimental values on the ordinate axis, while r'^2_m using them on the abscissa.
Note: r^2 is the same as R^2_{EXT} in formula 1). The different notations are kept for consistency with those reported in the literature[14,15].
6) $CCC = \frac{2\sum_{i=1}^{n_{EXT}} (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sum_{i=1}^{n_{EXT}} (y_i - \bar{y})^2 + \sum_{i=1}^{n_{EXT}} (\hat{y}_i - \bar{\hat{y}})^2 + n_{EXT}(\bar{y} - \bar{\hat{y}})^2}$ (ref. 16 and 17)
In this work we use the symbol y for the experimental data and $\hat{y}$ for the external predictions instead of x, as in the previous work (where we used x in accordance to the Lin's notation[17]). We made this change in order to unify the notations in this table. This formula, named now CCC (as commonly reported in the literature), was reported in our previous work as $\hat{\rho}_c$ in accordance with Lin's notation.
7) $RMSEP = \sqrt{\frac{\sum_{i=1}^{n_{EXT}} (y_i - \hat{y}_i)^2}{n_{EXT}}}$
8) $MAE = \frac{\sum_{i=1}^{n_{EXT}} \lvert y_i - \hat{y}_i \rvert}{n_{EXT}}$
9) $k = \frac{\sum_{i=1}^{n_{EXT}} y_i \hat{y}_i}{\sum_{i=1}^{n_{EXT}} \hat{y}_i^2}$, $k' = \frac{\sum_{i=1}^{n_{EXT}} y_i \hat{y}_i}{\sum_{i=1}^{n_{EXT}} y_i^2}$
TR = training set, EXT = external prediction set, y_i = experimental data values, $\hat{y}_i$ = predicted data values, $\bar{y}$ = average of the experimental data values, $\bar{\hat{y}}$ = average of the predicted data values.

in accepting a model as externally predictive for new chemicals, compared with the normally used cutoff values of the studied parameters. Table 1 shows the formulas of the compared validation criteria, along with Root Mean Square Error (RMSE), Mean Average Error (MAE), and the slopes of the regression lines (k and k'). They are reported to uniformly list all the reported parameters that are sometimes interchanged or not well calculated (here the CCC formula, previously named as $\hat{\rho}_c$ in accordance with Lin,[17] is harmonized in symbols with the other criteria). The r^2_m metric, compared in this second study, is not the one previously considered,[16] as it has now been updated in accordance with the last revision of Roy et al.,[14] where the mean value ($\overline{r^2_m}$) and the differences (Δr^2_m) were introduced.

The aim of this new work is to verify the performance of the above-reported external validation criteria in checking the agreement of predicted values with experimental ones in different extreme situations of deviation in data distribution. This is done by studying and comparing their general trends in depending on different possible biases (location or/plus scale shifts) in external data distributions, by means of a wide range of different simulated scenarios. Additionally, we want to verify and propose, on these comparative simulations, new recalibrated and intercomparable threshold values. This is done using the same level of data scattering for each validation criteria for the acceptance of good QSAR models as externally predictive, in a precautionary approach.

It is important to highlight that it is not mathematically correct to compare different validation criteria, which are based on very dissimilar formulas and consequently of very different behavior, using the same threshold value.[18] This point is particularly relevant as CCC can be considered a special, modified form of the correlation coefficient, and Q^2_{Fn} and r^2_m as sorts of determination coefficients (squared correlation coefficient); thus CCC is more or less comparable to the square root of the other criteria. To achieve our aim and compare all the criteria in the same scenarios, we specify an arbitrary scatter level of plotted data (with no bias), selected as acceptable, from our experience, for QSAR models. Using this level, we determine an intercomparable threshold for each criterion, then we use it to verify how well each coefficient detects the different kinds of biases (location, scale, location plus scale shifts).[17] This objective is achieved by simulations, not just by any specific QSAR model, which is not necessary for this kind of data agreement evaluation. The comparison has been also done for other scatter levels, obtaining qualitative similar results. Verification of the corresponding plots was always made in order to highlight, also visually, whether the criteria values are insensitive to some biases. If so, they could be

dx.doi.org/10.1021/ci300084j | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

prone to accept as externally predictive QSAR models, which do not make good external predictions.

No comparison is made in this paper of the Golbraikh and Tropsha (GT) method[1] because it is based on a set of metrics, thus, making it difficult to evaluate it in a single, representative value, as it is requested in this kind of work. However, this method is, in our opinion, one of the best for external validation. In addition, we previously demonstrated[16] it to be in good agreement with CCC.

We also wish to highlight the importance of a necessary contemporaneous visual inspection of scatter plots of experimental versus predicted data. In fact, as Pogliani and coauthors have already pointed out in their papers,[19,20] the plots can be nearly considered as the "spectra" of a model. However, scatter plots are not always reported in QSAR papers, and the reader is often obliged to go through papers, burdened with often huge tables reporting the corresponding statistics, usually only those preferred by the author. Thus referees and readers, without the plots to hand, would not notice possible problems, which could remain not evidenced by the apparent good values of the applied statistical criteria. On the contrary, experimental versus predicted value plots gives a visual easy check of the quality of a model and shows where, why, and to what degree the model fails; indeed, plots can confirm and strengthen statistical results: "*A picture is worth a thousand words*".[19]

In this context, a preliminary crucial point is that it is important to remember that even though the linear relationship between experimental and predicted data can be perfect (the coefficient of determination on external set $R^2_{EXT} = 1$ and also $R_0{}^2_{EXT} = 1$, i.e., the regression line passing through the origin; see Figure 1 that shows extreme theoretical examples not occurring in practice in good QSAR models), it does not automatically mean that the predicted data perfectly match the experimental ones (as already highlighted by Golbraikh and Tropsha[1]). In fact, the data match only if they lie on the diagonal of the experimental vs predicted data scatter plot. If, in published papers, R^2_{EXT} (sometimes called R^2_{pred}) is reported as the only measure of predictivity, and visual inspection of the corresponding scatter plot of experimental vs predicted data is not possible, the reader could be not completely confident on the real external predictivity of the model.

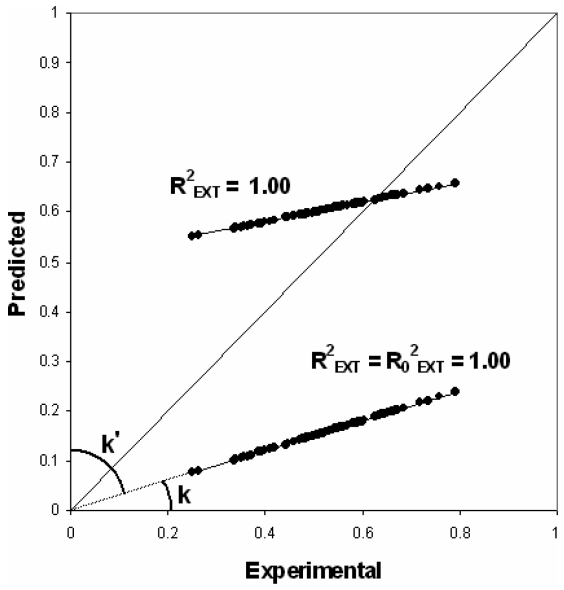


Figure 1. Examples of aligned external prediction data with R^2_{EXT} and/or $R_0{}^2_{EXT} = 1$ for not predictive models.

On the basis of previous comments, we confirm, as already stated in our previous work,[16] that in situations like the aligned data in the bottom part of Figure 1 (where $R^2_{EXT} = R_0{}^2_{EXT} = 1$), r^2_m failed and also $\overline{r^2_m}$ fails to detect the differences among experimental and predicted data when the regression line slopes (k and k') are not near to 1 (see Explanation SI1-1 of the Supporting Information for further details, where we demonstrate that $r_0{}^2$ and consequently $\overline{r^2_m}$ can be 1 and Δr^2_m can be 0, when the data are aligned, but the ordinate values are 10 times the ones on the abscissa, contrary to what was recently reiterated by Roy[18] and not by any demonstration). The closeness of the slopes k and k' to 1 is fundamental; in fact, it is a necessary condition for a predictive model, as already demonstrated by Golbraikh and Tropsha.[1]

METHODS

Generating Simulated Data Sets. The aim is to generate simulated prediction data sets for the comparison of experimental vs predicted data and upon which different levels of bias (location plus/or scale shift) can be applied.

The first step is to generate the random distributed data: here the Gaussian function $y = e^{-(x-X)^2/2\sigma^2}$ is applied using an arbitrary standard deviation (σ) of 0.15 and an average value (X) of 0.5. In order to generate the values, the first step is to extract a random number (x) from the data range of (0, 1). At this point we have to choose, according to the Gaussian function, whether this value must be kept or not. To accomplish this task, an additional random value (r) is extracted from 0 to 1 and compared to the y value of the Gaussian function calculated using the candidate value x. If the extracted random value (r) is smaller or equal to the y value, the candidate x value is accepted.

At this step, all data are aligned within the range of (0, 1) and are more concentrated in the middle. In this case, the data are scattered along just one axis. To build (when needed) a more realistic scenario for each just extracted value, an additional level of scattering can be added normally to the just cited axes. In this case, once the x value has been accepted, a new candidate random value r is extracted from −0.5 to +0.5, and a new value of y is calculated with average (X) equal to zero, using as standard deviation (σ) an arbitrary value called here "scattering" (in this work it ranges from 0.0025 to 0.06 using steps of 0.0025). If a new extracted random value ranging from 0 to 1 is smaller that the just calculated y value, the extracted value r (the random number from −0.5 to +0.5) is considered the normal "dispersion" of the main candidate value x. Graphically, we would obtain a broadly ellipsoid cloud of points with the major axis lying on the same axis where all the x data lay before scattering.

Once obtained the data, scattered or not, are rotated 45° and then shifted to make the centroid of the data matching the 0.5 coordinates of both the abscissa and the ordinate axes of the plot. Summarizing, we obtain a more or less scattered cloud of data points along the diagonal, where the centroid matches the coordinates (0.5, 0.5), i.e., the middle of the proposed range. We now arbitrarily keep the abscissa coordinates of every point as the experimental data and the corresponding ordinate coordinates as the predicted values (the fact that no underlying modeling has been assumed is explained in the Simulated Data Assumptions section).

dx.doi.org/10.1021/ci300084j | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

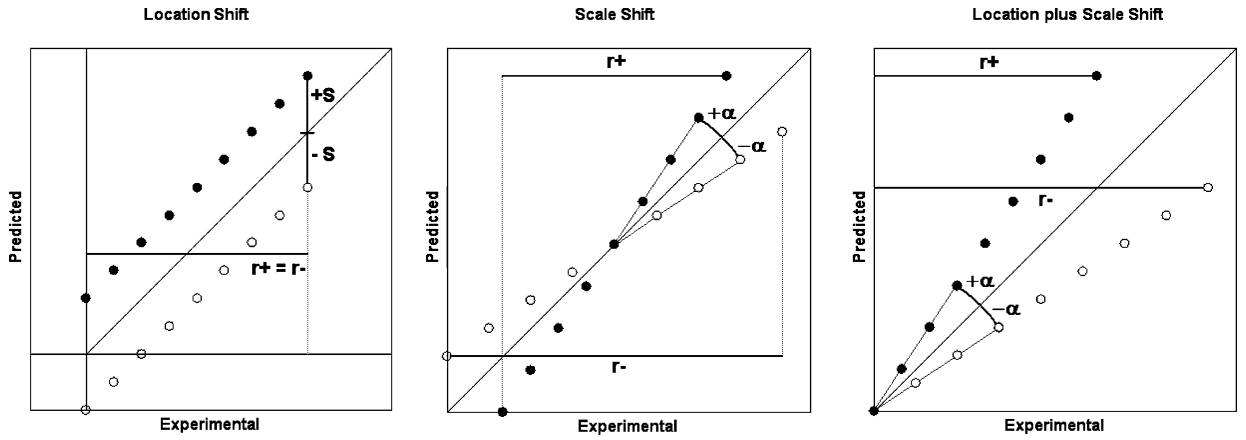


Figure 2. Deviations of the experimental values from the predicted ones. "S" is the shift of the predicted values from the diagonal (perfect agreement), "+" is the addition of a certain value, while "−" means a subtraction, "α" means a positive (+) or negative (−) rotation of the plotted data with respect to the diagonal. In the Scale Shift rotation, the pivot is in the middle of the data range on the diagonal, while in the Location plus Scale Shift is on the origin. Note that "r−" is the range of experimental responses corresponding to a decrease in the applied shift or angle, and "r+" is the range corresponding to an increase in the applied shift or angle.

Once the data are generated, centered to the range of bias and lying on the diagonal, the biases (location plus/or scale shift) are applied in the following manner: (1) Regarding the location shift, the data are shifted upward or downward following the ordinate axis ranging from −0.3 to 0.3 with steps of 0.0005. (2) Regarding the scale shift, the data are rotated from the diagonal using as a pivot the coordinate (0.5, 0.5) with a range of angles spanning from −30° to 30° using steps of 0.05°. (3) Regarding the scale plus location shift, the data are rotated using as a pivot the origin of the graph (0, 0) with the same range and angle steps from the diagonal reported in the previous point 2.

The aforementioned procedure, that implies 1201 simulated data distributions per type of bias, and for every step, is repeated 100 times using different random data, and then the whole procedure is repeated for every scattering level (25 levels), leading to a total of 9,007,500 generated external data sets. All the studied validation criteria were calculated for each level of scattering (i.e., from 0 to 0.06 with steps of 0.0025), but in this work, we focus only on those pertaining to the levels of 0 and 0.04 (all the unreported results of the remaining levels of scattering are qualitatively consistent). The first is used for the qualitative analysis (for the general behavior), while the latter is for the quantitative one (used for the determination of the new cutoffs).

Simulated Data Assumptions. From the formulas of the validation criteria in Table 1, it can be seen that the calculation of CCC, $\overline{r^2_m}$, and Q^2_{F2} does not need the training set, while that of Q^2_{F1} and Q^2_{F3} does. Indeed for CCC, $\overline{r^2_m}$, and Q^2_{F2}, there is no need to simulate a model because it is sufficient to plot the experimental vs predicted values. However Q^2_{F1} and Q^2_{F3} cannot be calculated just from the external data points because the $\bar{y}_{TR}$ value is needed, as well as all the training set experimental values for the Q^2_{F3} calculation. As we wish to focus only on the shift and/or rotation biases of the predicted data with respect to the experimental without assuming any underlying model, it follows that we can simulate scenarios where the "virtual" training set matches the prediction set. Even though this assumption may seem unrealistic, if one assumes that the training set range is similar to that of the prediction set, as in real QSAR studies, their averages are expected to be similar. In this study we accept for simplicity that the averages match, so the results are based on this reasonable approximation. However, another simulation protocol, not reported here, was run by generating separated training sets, corresponding models, and simulating the prediction sets. Such a more realistic simulation was used to cross-check the protocol reported here, obtaining very similar results. When the averages of the training and prediction set match, the highest possible values of Q^2_{F1} and Q^2_{F3} are obtained, so as a consequence, their corresponding cutoffs are the highest and are more restrictive in comparison to the values we could obtain when the averages differ. Therefore, in this case, we use those validation criteria in the best setup.

It is also important to note, looking at their formulas, that Q^2_{F1} and Q^2_{F3} should give different results, even though they use the same "virtual" training set. In fact, if the range of experimental responses ($r+/r-$ in Figure 2) changes (e.g., rotating the data plotting of the experimental responses vs the predicted with respect to the center of the graph), as depicted in the figure, the range of the experimental response values will differ. In fact, the denominator of Q^2_{F1} is computed using the prediction set values, where the range and/or average value changes, while in Q^2_{F3} these are fixed because of the assumptions pertaining to the training set, which is computed once when the data are generated (i.e., before shifting and/or scaling the responses). Because the simulation setup also assumes that the number of the elements in the training set equals the number in the prediction set, the Q^2_{F3} formula can be simplified as

$$Q^2_{F3} = 1 - \frac{\sum_{i=1}^{n_{EXT}} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n_{TR}} (y_i - \bar{y}_{TR})^2} \quad (1)$$

The calculation of $\bar{y}_{TR}$ using the proposed simulation setup leads to an important topic concerning Q^2_{F1} and Q^2_{F2}. In fact, it is expected that in each situation where $\bar{y}_{EXT}$ does not change, e.g., the location shift setup (Figure 2), the values of Q^2_{F1} and Q^2_{F2} will match. In addition, situations where $\bar{y}_{EXT}$ is expected to change by a relatively small degree, e.g., the location plus scale shift setup (Figure 2), the values of Q^2_{F1} and Q^2_{F2} are expected to be similar. This would lead to the wrong conclusion that Q^2_{F1} and Q^2_{F2} have basically the same performance, but in real situations, this usually does not happen as the values of $\bar{y}_{TR}$ and $\bar{y}_{EXT}$ can vary even to a consistent degree. The similarity of $\bar{y}_{TR}$ and $\bar{y}_{EXT}$ values, and consequently the similarity of the Q^2_{F1} and

dx.doi.org/10.1021/ci300084j | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

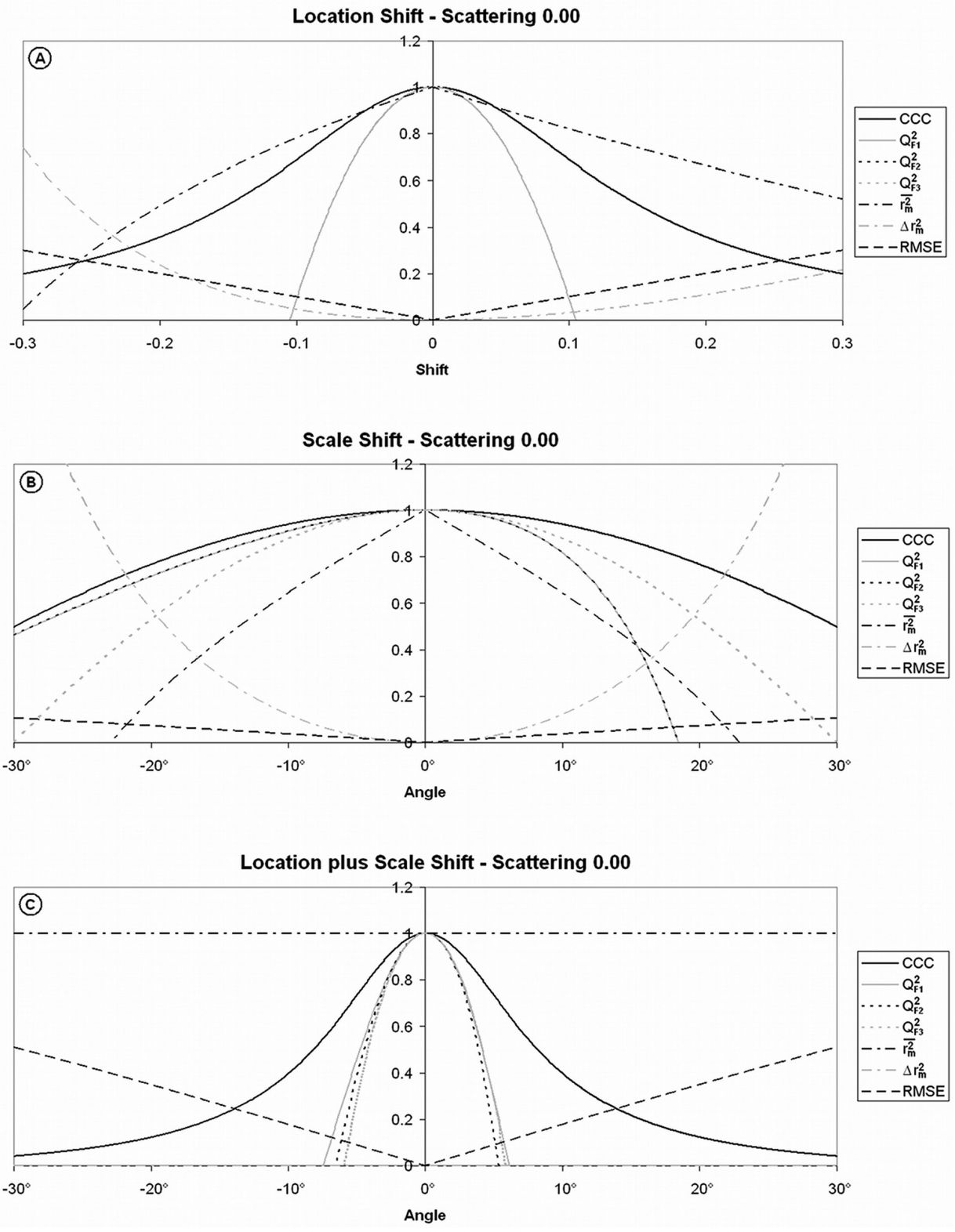


Figure 3. Validation criteria comparison (values on y axis, maximum value for all = 1) at different location and scale shift levels (x axis) using simulated experimental values vs prediction values with no scattering.

Q^2_{F2} values, is mainly a consequence of the assumptions made in this work. The values of Q^2_{F1} are calculated under ideal conditions, but in real situations, Q^2_{F1} is usually overoptimistic in comparison to Q^2_{F2},[10] and this work does not prove the contrary. In fact, in our previous work[16] where we used more realistic simulations (all the predicted data were generated by a model using different data set sizes and different training-prediction set proportions), Q^2_{F1} proved to be more optimistic than the other Q^2_{Fn} validation criteria, leading to a smaller proportion of agreement in accepting/rejecting models in certain setups.

With regard to Q^2_{F3}, the situation differs from that of Q^2_{F1} in that, as already stated, the values of y_i are taken from the

dx.doi.org/10.1021/ci300084j | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

training set and thus will not change on applying the different rotation/shift biases to the external responses.

Reference Validation Criteria Thresholds. In this work, as reference values for calculating new thresholds, we used the threshold 0.6 for Q^2_{Fn}. The threshold values of $\overline{r^2_m}$ (0.5) and Δr^2_m (0.2) are those suggested by the proponent.[14]

RESULTS AND DISCUSSION

On analyzing the possible impact of the deviation of the experimental values from the predicted ones on the different validation criteria, we adopted two approaches using simulated data: (i) a qualitative analysis of the behavior of the validation criteria, studying and understanding the reason for the possible trends simply from their formulas, and (ii) a quantitative analysis (see Methods section) to quantify the general trends of the validation criteria in the presence of different bias levels, to compare the respective cutoffs and to propose new ones, more intercomparable and precautionary.

Validation Criteria and Biases in the Prediction Set. The study of the impact of the possible deviations of the predicted values from the experimental ones, on the different validation criteria, is pivotal to this work. Predicted data can deviate from experimental values in many ways; thus, we study the three possible trends proposed by Lin,[17] here exemplified in Figure 2.

Here, the abscissa axis corresponds to the external experimental data, while the ordinate axis corresponds to the predictions for the external data set. Such a disposition is chosen because it is more usual in QSAR studies. However, it is important to note that some of the studied validation criteria (CCC, Q^2_{Fn}) have the additional advantage to be insensitive to axis disposition.

Location shift (see first graph in Figure 2) means that all the predicted responses are systematically shifted upward or downward from the diagonal (perfect agreement). In practical QSAR examples, this can happen when the model tends to over- or underestimate the experimental data. This situation could occur, for instance, if the external data were obtained from a different source that yielded overall higher endpoint values compared to the training data, thus introducing a systematic error. As can be noted for this kind of shift, the range (r+ and r−) of the experimental values is always the same.

Scale shift (see second graph in Figure 2) means that all the responses are rotated by a positive or negative angle from the diagonal with respect to the center of the range. It can be noted that in this kind of shift the range of the experimental responses changes, depending on the applied angle. If the angle increases with respect to the diagonal ($+\alpha$), the range decreases (r+), while it increases (r−) if the angle decreases ($-\alpha$).

Location plus scale shift is conceptually similar to scale shift, but the pivot of the rotation lies on the origin of the graph. In this case, the range of the experimental responses behaves qualitatively as in the scale shift setup.

The latter two kinds of biases are harder to exemplify in the QSAR practice in comparison to the location shift. However, we recall that this study is a theoretical one and must consider also extreme possibilities that anyway cannot be excluded *a priori*.

Validation Criteria Behavior. Figure 3 shows the trends of all the studied validation criteria in relation to the three different biases.

Location Shift Analysis. Figure 3A shows the graphs of all the considered validation criteria and the corresponding RMSE values at different values of location shift. At first glance, as expected, the bigger the location shift (on x axis) the poorer the agreement of the experimental values with predicted ones; thus, all the validation criteria decrease from the maximum value of 1. The first evident feature is that, in this case, all the Q^2_{Fn} coincide (see Explanation SI1-2 of the Supporting Information for further details). They respond to the location shift symmetrically and, of all the other validation parameters, are the most sensitive to increase or decrease in this kind of shift (as they fall down more rapidly). CCC, also symmetrical, is less sensitive to the increase or decrease of the location shift, as expected from the analytical form. Because, as already stated, CCC is more or less comparable to the square root of Q^2_{Fn}, it is expected that Q^2_{Fn} will decrease "sooner" than CCC, as the location shift increases or decreases.

On the contrary it is less understandable that $\overline{r^2_m}$, which like Q^2_{Fn} is similar to a determination coefficient, is less sensitive to the increase or decrease in the location shift and is asymmetrical with respect to zero shift (see Explanation SI1-3 of the Supporting Information for further details). The fact that $\overline{r^2_m}$ gives different results for the same differences among experimental vs predicted data (while the RMSE is symmetrical) and that it also tends to maintain high values of model acceptance in cases of high data shift is a warning to the QSAR developer using this validation criterion in cases of similar data shift, visible only after model development from the corresponding plot.

Scale Shift Analysis. Figure 3B shows the graphs of all the considered validation criteria values for different scale shift values and the corresponding RMSE lines. At first glance it can be noted that Q^2_{F1} and Q^2_{F2} coincide, and that they are asymmetrical with respect to the zero angle of applied scale shift (see Explanation SI1−4 for more details). For negative angles of the scale shift, they are similarly sensitive as CCC, while they are more sensitive in comparison to CCC for positive angles. This asymmetry with respect to zero suggests that, at least for scale

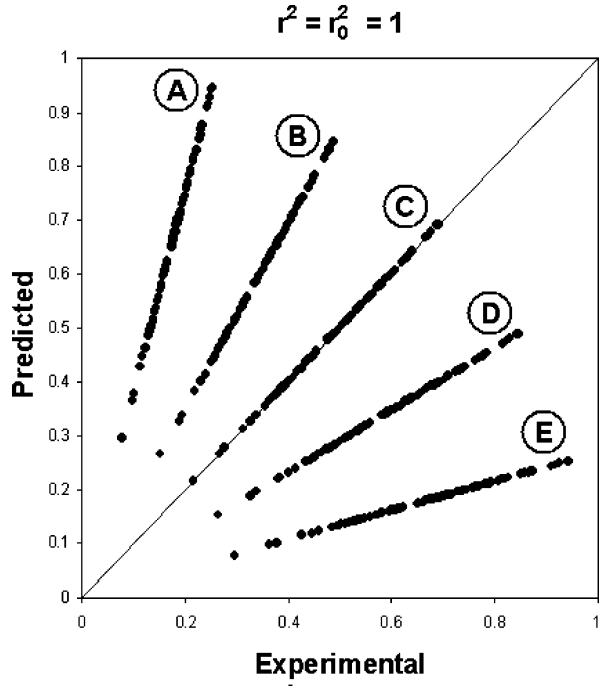


Figure 4. Cases where $\overline{r^2_m}$ is always 1 and Δr^2_m is always 0.

dx.doi.org/10.1021/ci300084j | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

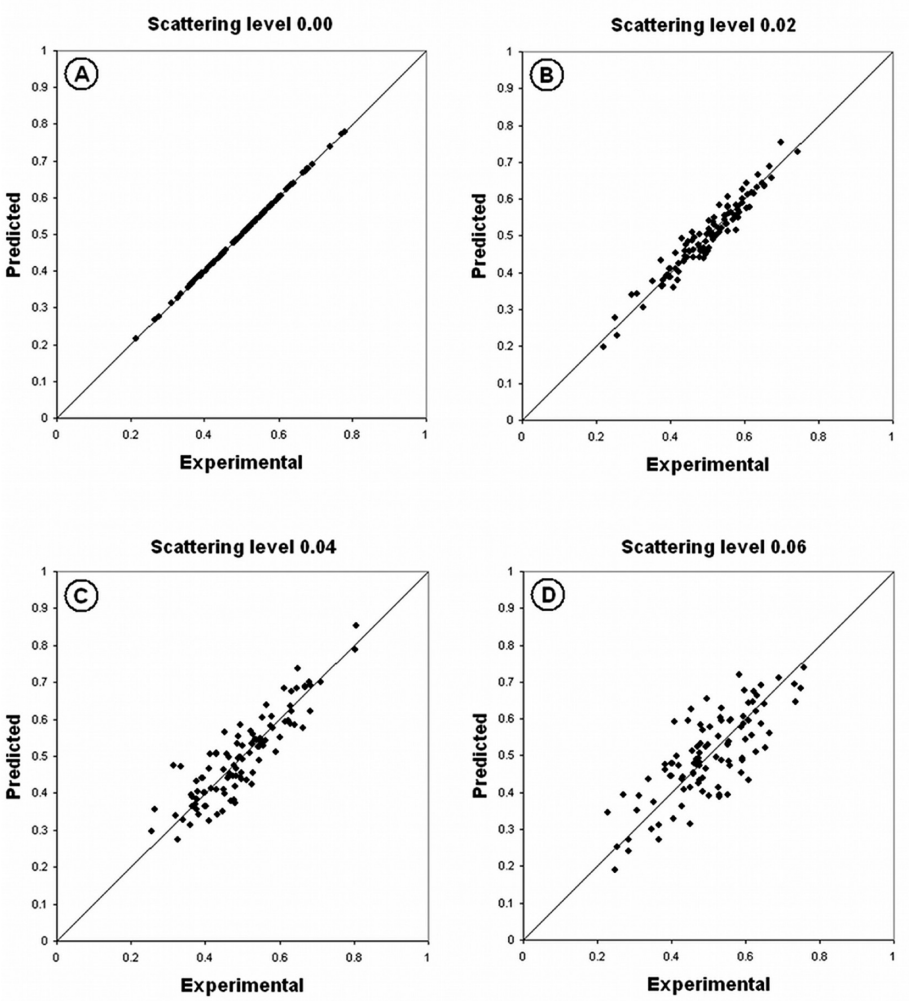


Figure 5. Levels of scattering in simulated data sets.

shift, Q^2_{F1} and Q^2_{F2} should be used with caution, evaluating the magnitude and the sign of the angle involved in real data sets. This behavior is unexpected and suspicious, mainly because this asymmetry does not reflect the symmetry of RMSE.

Q^2_{F3} is symmetrical (see Explanation SI1-4 of the Supporting Information for more details) with respect to zero like CCC but is more sensitive to increases or decreases in the applied angle. As expected, $\overline{r^2_m}$ is symmetrical because for the same absolute value of the rotation angle $\pm\alpha$, as represented in Figure 2, the swapping of the axes results in symmetrical graphs; in this particular case of data shift, it appears to be the most sensitive.

Location Plus Scale Shift Analysis. Figure 3C shows the graphs of all the considered validation criteria for different values of location added to scale shift and the corresponding RMSE lines. The first evident and serious anomaly concerns $\overline{r^2_m}$, which is always 1 (and $\Delta r^2_m = 0$ because $r^2 = r^2_0 = 1$ whatever the applied angle). This metric, which ignores the necessary condition for predictive models to have either k or k' near 1,[1] is not able to discriminate among the different applied angles for the location plus scale shift bias. As a consequence, on the basis of the $\overline{r^2_m}$ values, all the models with plots as in Figure 4 (where only case C can be considered as predictive, while A, B, D, and E cannot) could be accepted as externally predictive regardless of the applied shift angle, but such tolerance is certainly unacceptable in QSAR modeling.

Both CCC and Q^2_{F3} are symmetrical, while Q^2_{F1} and Q^2_{F2} are slightly asymmetrical with respect to zero (see Explanation SI1-5 of the Supporting Information).

All the Q^2_{Fn} graphs are "steeper" than CCC, "quickly" falling to low values as the applied angle increase or decreases and, without considering the slight asymmetry of Q^2_{F1} and Q^2_{F2}, the scenario is similar to that of the location shift analysis. Thus, comparable reasoning to that above on sensitivity can be applied.

Levels of Scattering. In the following quantitative analysis, in order to determine the maximum allowable scattering level, i.e., the value acceptable to determine the validation criteria thresholds, different levels of external data scattering are generated (see Methods section) and plotted as in the examples in Figure 5.

In previous paragraphs, we applied the values of shift to external data points lying along the same line (Figure 5, scattering level 0.00). However, such a scenario is unrealistic in QSAR modeling because of the lack of data point dispersion away from the line.

Even though, as specified in the Methods section, we evaluated every scattering level from 0 to 0.06, our experience in QSAR modeling and an analysis of the Figure 5 graphs led us to arbitrarily select the 0.04 scattering level as the highest acceptable level for satisfactorily predictive QSAR models. Thus, this arbitrary level was chosen to determine the validation criteria thresholds for reliable QSAR models in a precautionary approach.

dx.doi.org/10.1021/ci300084j | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

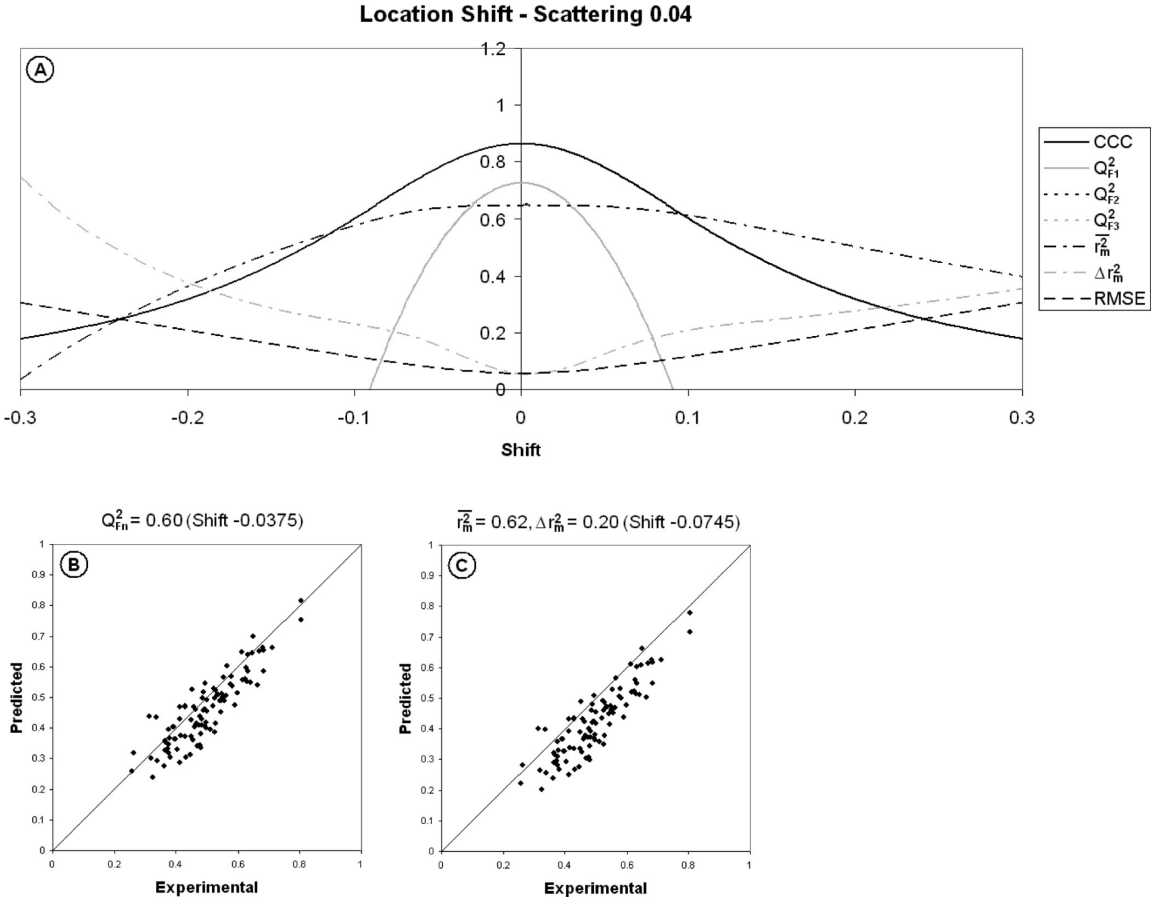


Figure 6. Validation criteria comparison at different location shift levels using simulated experimental values vs prediction values with a scattering level of 0.04. Some explicative examples of the external data plotting are reported. For the model in B: CCC = 0.81, Q^2_{Fn} = 0.60, $\overline{r^2_m}$ = 0.65, and Δr^2_m = 0.12. For the model in C: CCC = 0.70, Q^2_{Fn} = 0.24, $\overline{r^2_m}$ = 0.62, and Δr^2_m = 0.20.

Concerning RMSE and CCC. Looking at the reported simulations and those used in our previous work,[16] it can be noted, as expected, that the smaller the CCC the higher the RMSE value. But there are situations where for the same value of RMSE, the CCC varies. To exemplify, if the plotting of Figure 5C is extended along the diagonal adding more points maintaining a comparable scattering level, the RMSE value would be approximately the same, while the CCC increases. The CCC value,[21,22] as well as all the here-studied validation criteria (except probably Q^2_{F3},[12]) depends on the range of the external data. This dependence could be considered as a disadvantage, but this point is questionable. In fact, given a certain data scattering level, a fixed threshold would help to verify if the external data range is sufficiently wide. For example, in cases where the scattering level could be considered acceptably "small" (but remember that RMSE depends on the scale of measure), small validation criteria values could be obtained if the range is not wide enough. In these cases, the QSAR developer is probably overrating the prediction ability of the model under analysis.

Validation Criteria Threshold Determination. Figures 6−8 show the trends of all the studied validation criteria in relation to the three different biases at the chosen 0.04 scattering level. Here, we remember that the values of Q^2_{F1} are calculated under ideal conditions (see Simulated Data Assumptions). However, in real situations, Q^2_{F1} is usually overoptimistic in comparison to Q^2_{F2},[10] and this work does not prove the contrary. Table 2 allows the comparison of all the values of the validation criteria in relation to the various shifts.

Location Shift Analysis. From Figure 6A, it can be seen that the general behavior of all the criteria is qualitatively comparable to that of corresponding Figure 3A. As expected, for zero shift, the maximum value of the validation criteria (Q^2_{Fn}, CCC, and $\overline{r^2_m}$), derived from the simulation exercise at 0.04 scattering, no longer corresponds to 1 (and RMSE is no longer zero). In addition, they all differ: CCC is 0.86, Q^2_{Fn} values are 0.72, $\overline{r^2_m}$ is 0.65 (and Δr^2_m = 0.06, thus acceptable as Ojha et al. proposed).[14] With these cutoff values, all the studied validation criteria are consistent in accepting as predictive QSAR models with a plotting similar to that in Figure 5C (scattering level 0.04). Here, $\overline{r^2_m}$ is not very responsive to the change in shift, particularly for positive values, and its line is even "flatter" than in Figure 3A.

Taking into account the Q^2_{Fn} threshold of 0.6, which we considered in our previous paper,[16] and looking at the corresponding plots (for example, Figure 6B and Figure SI1-2A of the Supporting Information), a QSAR developer would probably reject the models because the predicted responses are shifted too downward or upward. The corresponding CCC value is 0.81 (thus, below the 0.85 threshold suggested in our previous work[16]), so in this case, we would suggest model rejection, while on the contrary, $\overline{r^2_m}$ would accept it (in fact, $\overline{r^2_m}$ = 0.65 and Δr^2_m = 0.11, greater than 0.5 and smaller than 0.2, respectively, of the cutoff values suggested by the Roy group[14]).

dx.doi.org/10.1021/ci300084j | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 2. Detected Thresholds for Each Systematic Shift Type, Using a Scattering Level of 0.04[a]

Shift type	Shift Angle	Fig.	CCC	Q^2_{F1}	Q^2_{F2}	Q^2_{F3}	$\overline{r^2_m}$	Δr^2_m
Location Shift	0	5c	**0.86** ± 0.03	**0.72** ± 0.05	**0.72** ± 0.05	**0.72** ± 0.05	**0.65** ± 0.06	**0.05** ± 0.04
Location Shift	-0.0375	6b	**0.81** ± 0.03	**0.60** ± 0.07	**0.60** ± 0.07	**0.60** ± 0.07	**0.65** ± 0.06	**0.12** ± 0.06
Location Shift	0.0375	SI2a	**0.81** ± 0.03	**0.60** ± 0.07	**0.60** ± 0.07	**0.60** ± 0.07	**0.64** ± 0.06	**0.10** ± 0.06
Location Shift	-0.0745	6c	**0.70** ± 0.04	**0.24** ± 0.12	**0.24** ± 0.12	**0.24** ± 0.12	**0.62** ± 0.05	**0.20** ± 0.04
Location Shift	0.0935	SI2b	**0.63** ± 0.04	**-0.04** ± 0.16	**-0.04** ± 0.16	**-0.04** ± 0.16	**0.62** ± 0.05	**0.20** ± 0.04
Scale Shift	0°	5c	**0.86** ± 0.03	**0.72** ± 0.05	**0.72** ± 0.05	**0.72** ± 0.05	**0.65** ± 0.06	**0.05** ± 0.04
Scale Shift	-18.30°	7b	**0.70** ± 0.04	**0.60** ± 0.04	**0.60** ± 0.04	**0.39** ± 0.05	**0.28** ± 0.05	**0.44** ± 0.04
Scale Shift	6.35°	7c	**0.84** ± 0.03	**0.60** ± 0.07	**0.60** ± 0.07	**0.68** ± 0.06	**0.59** ± 0.05	**0.22** ± 0.04
Scale Shift	-11.20°	7d	**0.80** ± 0.03	**0.70** ± 0.04	**0.70** ± 0.04	**0.60** ± 0.05	**0.48** ± 0.05	**0.29** ± 0.03
Scale Shift	10.55°	SI3b	**0.80** ± 0.03	**0.42** ± 0.10	**0.42** ± 0.10	**0.60** ± 0.07	**0.48** ± 0.05	**0.29** ± 0.03
Scale Shift	-5.65°	7e	**0.85** ± 0.03	**0.74** ± 0.05	**0.74** ± 0.05	**0.69** ± 0.05	**0.61** ± 0.05	**0.20** ± 0.04
Scale Shift	5.20°	SI3a	**0.85** ± 0.03	**0.63** ± 0.07	**0.63** ± 0.07	**0.69** ± 0.06	**0.61** ± 0.05	**0.20** ± 0.04
Location + Scale Shift	0°	5c	**0.86** ± 0.03	**0.72** ± 0.05	**0.72** ± 0.05	**0.72** ± 0.05	**0.65** ± 0.06	**0.06** ± 0.04
Location + Scale Shift	-2.50°	8c	**0.80** ± 0.03	**0.60** ± 0.06	**0.58** ± 0.07	**0.55** ± 0.07	**0.65** ± 0.06	**0.06** ± 0.04
Location + Scale Shift	2.00°	SI4f	**0.82** ± 0.03	**0.60** ± 0.07	**0.59** ± 0.07	**0.61** ± 0.07	**0.65** ± 0.06	**0.05** ± 0.04
Location + Scale Shift	-2.35°	SI4b	**0.80** ± 0.03	**0.61** ± 0.06	**0.60** ± 0.07	**0.57** ± 0.07	**0.65** ± 0.06	**0.06** ± 0.04
Location + Scale Shift	1.90°	SI4c	**0.82** ± 0.03	**0.61** ± 0.07	**0.60** ± 0.07	**0.62** ± 0.07	**0.65** ± 0.06	**0.05** ± 0.04
Location + Scale Shift	-2.15°	SI4d	**0.81** ± 0.03	**0.63** ± 0.06	**0.62** ± 0.06	**0.60** ± 0.07	**0.65** ± 0.06	**0.06** ± 0.04
Location + Scale Shift	2.10°	SI4e	**0.81** ± 0.03	**0.59** ± 0.07	**0.57** ± 0.07	**0.60** ± 0.07	**0.65** ± 0.06	**0.05** ± 0.04
Location + Scale Shift	-20.45°	8b	**0.11** ± 0.02	**-2.37** ± 0.22	**-6.27** ± 1.07	**-10.4** ± 1.6	**0.50** ± 0.07	**0.18** ± 0.06
Location + Scale Shift	20.10°	SI4a	**0.11** ± 0.02	**-1.74** ± 0.04	**-24.2** ± 3.4	**-10.1** ± 1.6	**0.50** ± 0.07	**0.15** ± 0.07

[a]Values reported for the validation criteria calculated over the simulated data are the averages ± standard deviations. Grey cells: fixed values from which all the others values on the same row have been detected. The reported figures (Figure column) refer to one example of the 100 used for every applied value of the shift/angle in the simulations.

Let us consider, as an extreme example, a model with a plot like that in Figure 6C (and similarly in Figure SI1-2B of the Supporting Information). A QSAR developer would reject similar models just by looking at the plots of the experimental vs predicted data. However, these models would be considered predictive only by the Roy's parameters with $\overline{r^2_m}$ = 0.62 and Δr^2_m = 0.20 (borderline value).[14]

Summarizing, as the Q^2_{Fn} threshold of 0.60 puts the QSAR developer into an ambiguous zone where it is not easy to decide whether to accept or reject a model, we suggest raising the Q^2_{Fn} threshold to 0.70 (a value that seems to be a good candidate, rounding the 0.72 value reported in Table 2) and the $\overline{r^2_m}$ threshold from 0.50 to a more restrictive value of 0.65 for the acceptance of good QSAR models. The suggested cutoff value of 0.85 for CCC, proposed in our previous work,[16] is confirmed here (rounding the 0.86 value reported in Table 2).

Scale Shift Analysis. On plotting the simulated data with different scale shift angles at the fixed scattering level of 0.04 (see Figure 7A), it can be seen that the general behavior is qualitatively comparable to that of the corresponding Figure 3B.

The same reasoning explained in the previous section for location shift, concerning the maximum values of the validation criteria, still applies, but it can be noted that the maximum values of Q^2_{F1} and Q^2_{F2} (that are coincident) are not at angle zero (see Explanation SI1-6 of the Supporting Information for further details). Therefore, the highest values of Q^2_{F1} and Q^2_{F2} do not correspond to the best external data plot and, as a consequence, to the smallest RMSE. This is not intuitive, but it must be noted that the lack of correlation between $Q^2_{F1,2}$ and RMSE has already been remarked upon in a previous work of Consonni et al.[11,12] Q^2_{F1} and Q^2_{F2} are differently sensitive to the scale shift in the plot of experimental vs predicted data; in fact, sensitivity changes a lot, depending on the sign of the applied angle. These two criteria are less able to detect the differences among the rotations, if the data are negatively rotated from the diagonal; in fact, the left line is much less "steep". Thus, their use as external validation criteria is doubtful in cases of biases of this kind, as evidenced by the corresponding plots. In fact, keeping Q^2_{F1} and Q^2_{F2} at the threshold of 0.6, it can be noted that the corresponding absolute values of the angles are very different (Figure 7B, angle = −18.30° and Figure 7C, angle = 6.35°).

Looking at the trends of the respective lines in Figure 7A, $\overline{r^2_m}$ appears to be the most restrictive criterion, limited to the QSAR models when the bias in the external data is the scale shift type with negative angles, while for positive angles, it depends on the applied thresholds, but remains the most restrictive for reasonable values of the angle.

Looking at Figure 7A, it can be seen that a not acceptably high value of Δr^2_m corresponds to a threshold value of $\overline{r^2_m}$ = 0.5. In order to find the least acceptable value of $\overline{r^2_m}$, keeping Δr^2_m fixed at 0.20, we verified that the corresponding value of $\overline{r^2_m}$ is 0.61 (see Figure 7D and Figure SI1-3A of the Supporting Information), as also reported in Table 2. In this case, for homogeneity with the previous section on the location shift, we suggest to raise this threshold to 0.65.

For this kind of shift, CCC appears less responsive to the applied bias, in comparison to the location shift and also to the other criteria, except Q^2_{F3} that has a not too dissimilar trend. The threshold of 0.6 for Q^2_{F3} allows more bias in comparison to CCC (see Figure 7E and Figure SI1-3B of the Supporting Information), so this threshold seems too permissive.

With regard to the thresholds and rounding the values reported in Table 2, the ones proposed in the previous section can be confirmed also for this kind of bias, i.e., Q^2_{Fn} = 0.70, $\overline{r^2_m}$ = 0.65, and CCC = 0.85.

dx.doi.org/10.1021/ci300084j | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

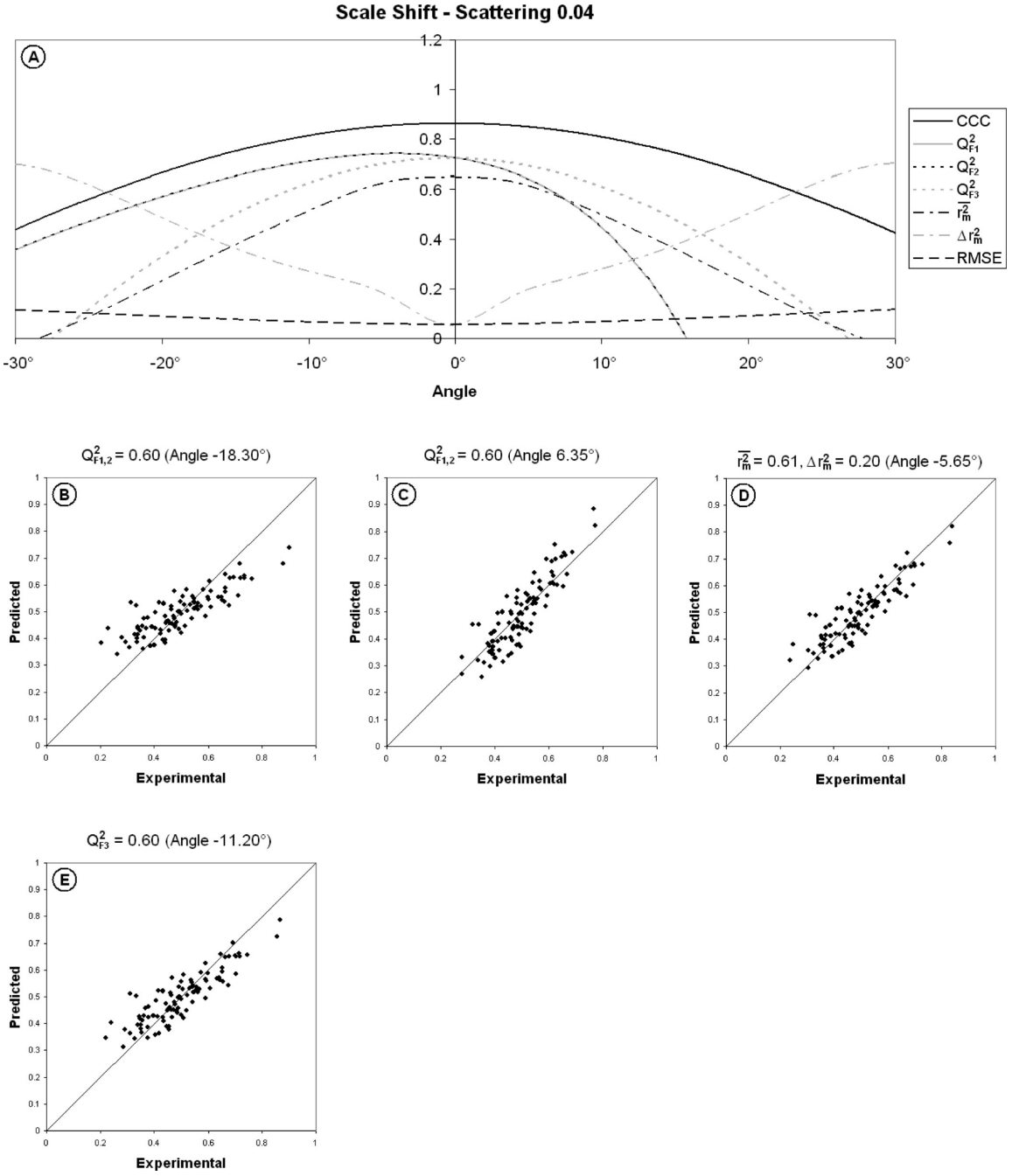


Figure 7. Validation criteria comparison at different scale shift levels using simulated experimental values vs prediction values with a scattering level of 0.04. Some explicative examples of the external data plotting are reported. For the model in B: CCC = 0.70, $Q^2_{F1,2}$ = 0.60, Q^2_{F3} = 0.39, $\overline{r^2_m}$ = 0.28, and Δr^2_m = 0.44. For the model in C: CCC = 0.84, $Q^2_{F1,2}$ = 0.60, Q^2_{F3} = 0.68, $\overline{r^2_m}$ = 0.59, and Δr^2_m = 0.22. For the model in D: CCC = 0.85, $Q^2_{F1,2}$ = 0.74, Q^2_{F3} = 0.69, $\overline{r^2_m}$ = 0.61, and Δr^2_m = 0.20. For the model in E: CCC = 0.80, $Q^2_{F1,2}$ = 0.70, Q^2_{F3} = 0.60, $\overline{r^2_m}$ = 0.48, and Δr^2_m = 0.29.

Location Plus Scale Shift Analysis. The shapes of the graphs in Figure 8A are qualitatively comparable to those of corresponding Figure 3C, so the same comments on symmetries and trends can be applied here. The value for $\overline{r^2_m}$ is slightly more responsive to change in the angles, i.e., it is not fixed at value 1, as previously shown in Figure 3C, but its line is still too "flat". Because of this very limited sensitivity to the increasing bias, $\overline{r^2_m}$ should be used with great caution in similar cases. With regard to the proposed threshold of 0.5[14] it is easy to verify that a corresponding QSAR model can result in highly biased graphs (see Figure 8B and Figure SI1-4A of the Supporting Information plots of models that would be accepted as predictive according to Ojha et al.[14] because $\overline{r^2_m}$ = 0.50 and Δr^2_m = 0.18). It is evident that this cutoff is highly misleading and could produce erroneous conclusions regarding QSAR model predictivity.

Using a threshold of 0.6 for Q^2_{Fn}, comparable plots of the three validation criteria are obtained, but similar kinds of bias (see Figure 8C and Figures SI1-4B to SI1-4F of the Supporting Information) confirm the previous sections' suggestion of the

dx.doi.org/10.1021/ci300084j | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

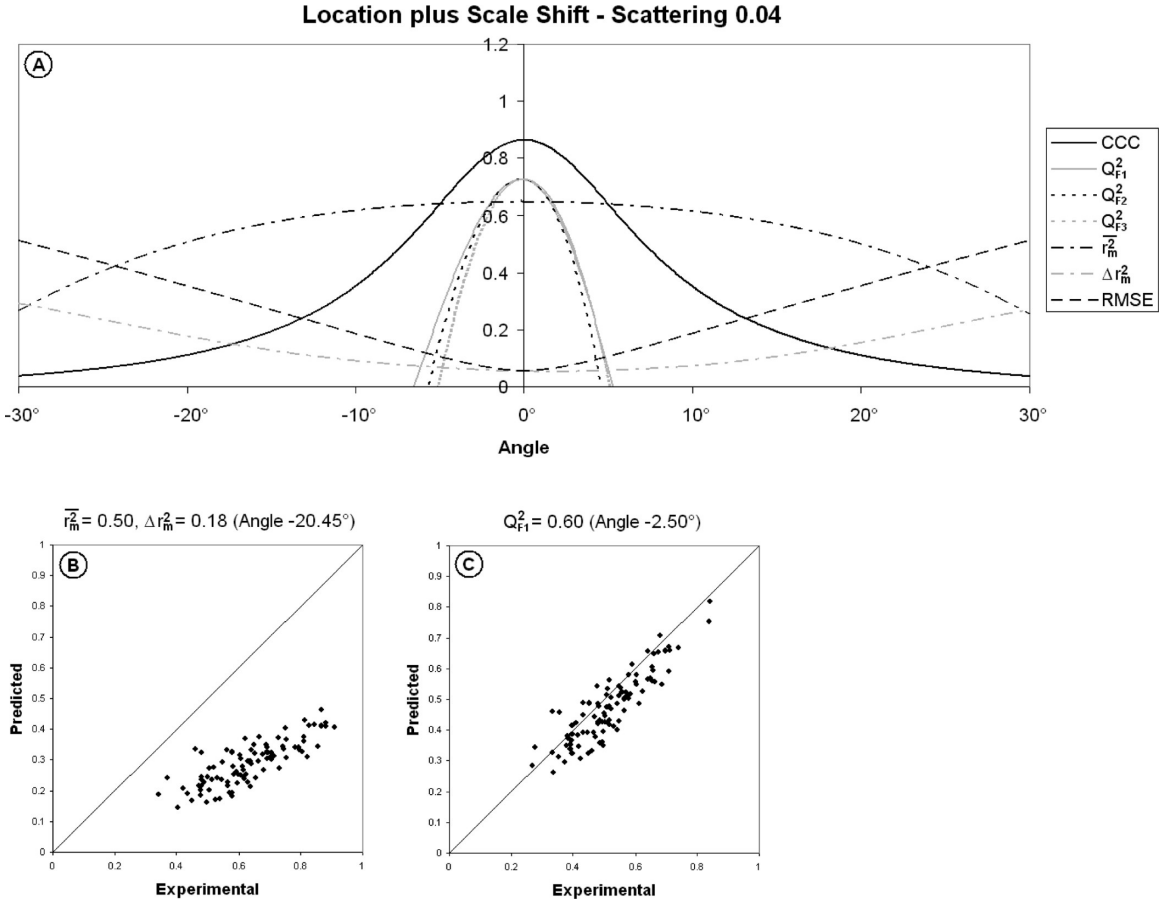


Figure 8. Validation criteria comparison at a different location added to scale shift levels using simulated experimental values vs prediction values with a scattering level of 0.04. Some explicative examples of the external data plottings are here reported. For the model in B: CCC = 0.11, Q^2_{F1} = −2.37, Q^2_{F2} = −6.27, Q^2_{F3} = −10.4, $\overline{r^2_m}$ = 0.50, and Δr^2_m = 0.18. For the model in C: CCC = 0.80, Q^2_{F1} = 0.60, Q^2_{F2} = 0.58, Q^2_{F3} = 0.55, $\overline{r^2_m}$ = 0.65, and Δr^2_m = 0.06.

need to raise the thresholds, if a more precautionary approach is preferred.

In summary, the thresholds we have proposed in the previous sections are confirmed here (Q^2_{Fn} = 0.70, $\overline{r^2_m}$ = 0.65, and CCC = 0.85), but in this case, we warn against the use of $\overline{r^2_m}$, as it is strongly insensitive to this kind of bias in the data, highlighted in the corresponding plots.

Once again, we wish to draw the QSAR developers' and also the readers' attention to the fact that if the graphs of the experimental vs predicted data are not looked at and just the validation parameter values are relied upon, unreliable models would be accepted as predictive. This highlights how important it is the contemporaneous use of scatter plots and validation criteria values.

Proposal for the Validation Criteria Thresholds. This section summarizes the results of the previous three sections (the "realistic" ones, where the scattering level is 0.04) and looks further at the corresponding proposed new precautionary and intercomparable thresholds for the validation criteria (Table 2).

The rows reported in Table 2, where some validation criteria are kept as reference values, based on previously proposed thresholds (grey cells), report the corresponding values of the remaining criteria. Thus, the aim of the full table is to give a general picture of the previously proposed acceptance values of Q^2_{Fn} and $\overline{r^2_m}$, suggesting that these values should be raised in a more precautionary QSAR modeling approach to the following values

$$Q^2_{Fn} = 0.70$$

$$\overline{r_m}^2 = 0.65$$

$$CCC = 0.85$$

Practical Application to QSAR Modeling: A Case Study. As a case study for practical application to QSAR models of the here-studied validation criteria with the intercomparable cutoff values, we selected a recent paper of Roy et al.[18] It reports comparative studies of the same validation criteria, based on a specific example of QSAR models developed on a single data set of the adsorption capacity of 3483 organic compounds to activated carbon in gas phase, a data set previously modeled by the Lanzhou University group in collaboration with Gramatica.[23]

Three models were developed on 2000 chemicals on the basis of three different splittings of the complete data sets (on structural similarity by cluster analysis (mod 1), on sorted responses (mod 2), and random (mod 3)). The corresponding models were then verified for their external prediction potential on 50 different sets (from 100 to 500 chemicals). There, all the validation criteria studied also in our present paper are

dx.doi.org/10.1021/ci300084j | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

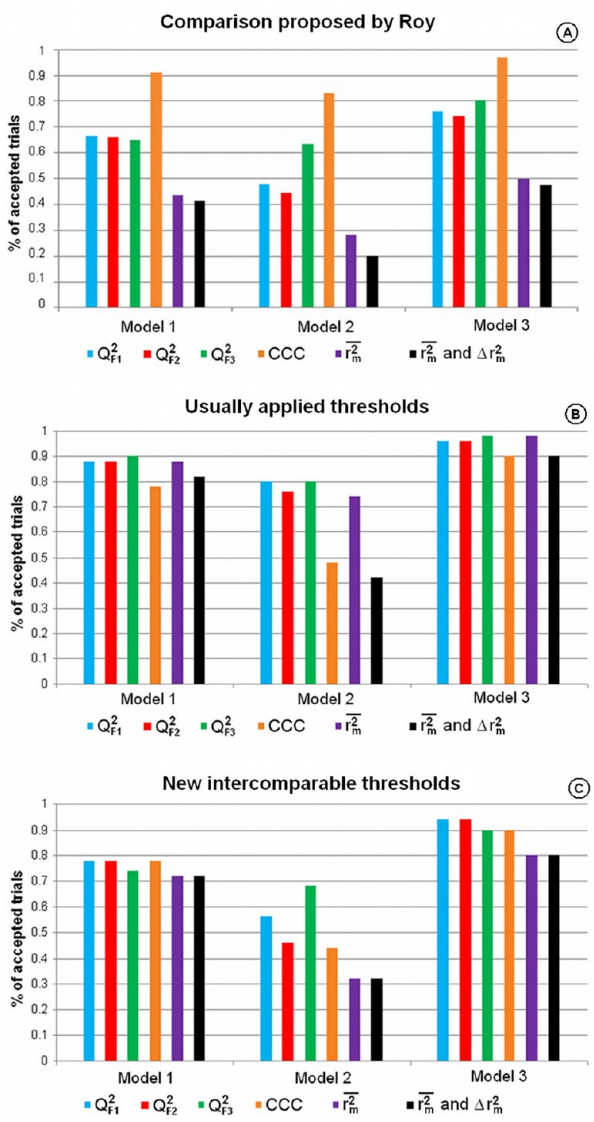


Figure 9. Percentage of models accepted by the considered validation criteria in the simulation proposed by Roy et al., using in A the usually applied thresholds (Q^2_{Fn} = 0.6, CCC = 0.85, $\overline{r^2_m}$ = 0.5, and Δr^2_m < 0.2), and in B, the new intercomparable and more precautionary proposed ones (Q^2_{Fn} = 0.72, CCC = 0.86, $\overline{r^2_m}$ = 0.65, and Δr^2_m < 0.2).

compared using the same threshold values for all the criteria, repeating the calculations by varying the thresholds (0.5, 0.6, 0.7, 0.8, and 0.85) on a total of 150 external sets. The Roy conclusion, highlighted also in their TOC graph (reported here in Figure 9A) is that "*CCC appears to be an overoptimistic metric, while* $\overline{r^2_m}$ *is the most conservative ...the most stringent criterion of external validation*".

In relation to this conclusion, some crucial points must be highlighted and commented on here:

On the basis of our previous demonstrations and those below, it is important to note that it is wrong to compare the different validation criteria using the same threshold values (as in Figure 9A). In fact, each criterion has different behavior, as we demonstrated previously (and in the present paper). This is a particularly important point for CCC, compared to other parameters, given that CCC is more or less comparable to the square root of the other criteria.

Therefore, the TOC graph of the Roy paper titled "*Comparison of stringent behavior of different external validation metrics*" should correctly be substituted by Figure 9B and C, where the results of the Roy modeling exercise have been recalculated and reported using the old cutoffs (as in our first paper[16]) and the new intercomparable and more precautionary ones proposed in this second paper.

From Figure 9, it is clearly evident that mainly for validation of models 1 and 3 the various criteria are in substantial agreement, and also in the third situation C, where the here-proposed new intercomparable cutoffs are used (making the $\overline{r^2_m}$ metrics even more restrictive in comparison to the cutoff proposed by Roy). However, the slightly higher restrictiveness of $\overline{r^2_m}$, along with Δr^2_m, is related only to three to five very similar external prediction sets (less than 10% of cases) (see plots of Figure 10A as an example and Figure SI2-1 of the Supporting Information, where all five trials in which the QSAR models were accepted as predictive by all the validation criteria, but rejected only by the r^2_m metrics using the new more restrictive cutoff values, are plotted). It is important to note that looking at the scatter plots these are peculiar cases, where we have verified that the predominant bias is the scale shift (note the angles in each graph). This particular kind of bias is the only one where the r^2_m metric is really the most restrictive criterion and contemporaneously CCC is the less stringent. This has been clearly demonstrated in our present paper (Figure 7).

Different comments can be made for the validation of Model 2 (based on sorted by response splitting), where in general less acceptable and discrepant results are obtained using all the criteria; in this case, it is apparent from the corresponding plots that in the majority of the models where the acceptance by the various criteria is different, the data are clustered in a few blocks (see the example in Figure 10B and others in Figure SI2-2 of the Supporting Information), probably because in this specific data set there are several blocks of very similar values in the responses, and this kind of splitting gives rise to a not uniform distribution of the data. In this peculiar case, the r^2_m metric shows it is better ability to recognize similar data distributions (certainly not good for QSAR modeling).

In conclusion, we verified on the specific example of Roy et al. models[18] that there is substantial agreement among the criteria if they are correctly compared using balanced cutoffs, both the old and the new intercomparable. The latest are additionally more precautionary, and CCC does not appear to be an overoptimistic parameter. In this specific modeling example, the r^2_m metric appears the most conservative because the models correspond to a scale shift, where r^2_m is indeed the most stringent criterion. However, its performance in the above specific case of bias in the data (scale shift) is a consequence of the modeling of that single particular data set, but it certainly cannot be considered as a generalizable behavior.

With regard to the slopes of the regression lines (k and k'), it is demonstrated here, in both analytical and visual form (see Figure 11 and Explanation SI1-1 and Figures SI1-1 and SI1-4A of the Supporting Information), that the existence of a linear relationship also passing through the origin (high values of R^2 and R_0^2) is a necessary but not sufficient condition for predictivity, as already highlighted by Golbraikh and Tropsha.[1]

The redundancy of the regression line slope, k and k', underlined in the Roy paper[18] on the basis of his specific example, is in our opinion a consequence of the particular data set used and the quality of the studied models. Normally, in QSAR studies, only

dx.doi.org/10.1021/ci300084j | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

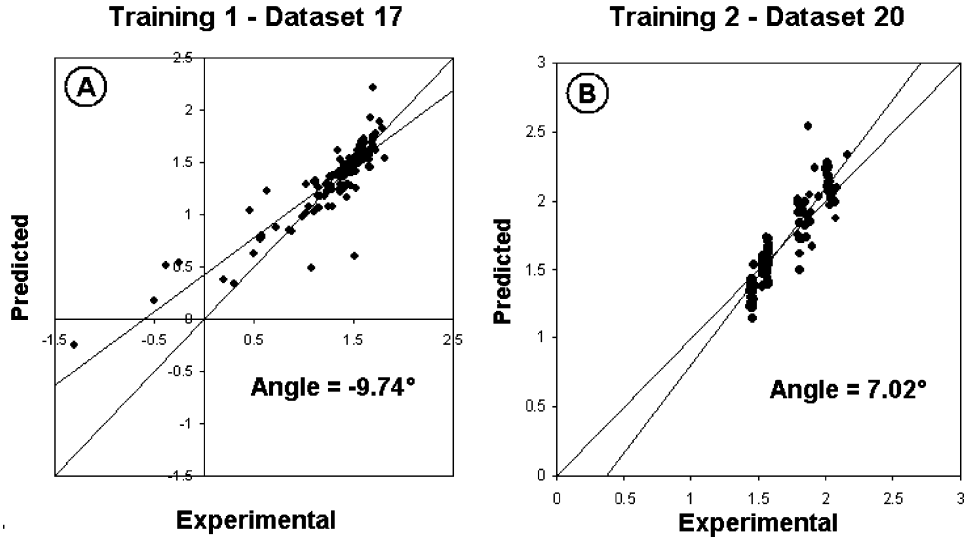


Figure 10. A: Example of data plot taken from the Roy work (Model 1 validation), with a scale shift, reported as the angle given by the difference between the diagonal and the regression line. B: Example of data plot taken from the Roy work with clustered data (Model 2 validation) and a scale shift.

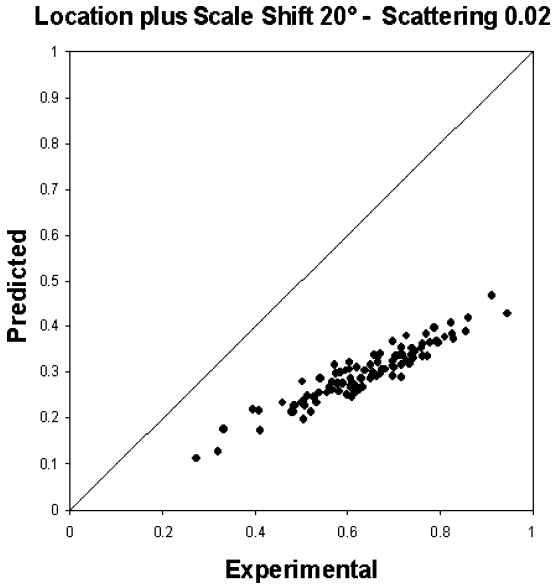


Figure 11. Example of simulated data plot where the r^2_m metric is overoptimistic because the slopes values (k = 2.135, k' = 0.466) are not near 1. Validation criteria values are: CCC = 0.11, Q^2_{F1} = −2.42, Q^2_{F2} = −6.71, Q^2_{F3} = −11.6, $\overline{r^2_m}$ = 0.84, and Δr^2_m = 0.10.

models with good performances in fitting and CV (high R^2 and Q^2_{LOO} values) are externally validated and because of the good internal quality of the tested model, it can be expected (and "dreamed") that the predictions are satisfactory; thus, in the scatter plot of the experimental vs predicted values, there will be closeness to the diagonal line (the slopes k or k' are near 1).

However, we verified by our simulations (Figures SI2-3 of the Supporting Information) that k and k' are not redundant in all scenarios, but they do appear superfluous or not useful (both being near 1: within 0.85−1.15, thus according to the GT criterion) only when a scale shift bias is present (Figure SI2-3B and C of the Supporting Information) i.e., exactly the bias we found in the Roy's specific example.[18] On the contrary, in cases of location or location plus scale shifts (Figure SI2-3D and E of the Supporting Information), k and k' are surely not redundant: only r^2_m accepts similar models, which does not guarantee the necessary condition of at least either k or k' near 1.[1]

Concluding this point, we confirm here the general relevance of the regression line slopes in checking the real external predictivity of QSAR models and also the need to look at the scatter plots of experimental versus predicted data. More detailed and additional comments on the parallel Roy paper are reported in Supporting Information SI3.

CONCLUSIONS

This work, following our previous proposal of using a conceptually simple parameter, the Concordance Correlation Coefficient (CCC) for external validation of QSAR models, studies the sensitiveness of the most commonly used validation criteria (Q^2_{F1}, Q^2_{F2}, Q^2_{F3}, $\overline{r^2_m}$, and CCC) to different kinds of bias in data (location or/plus scale shifts, as measures of accuracy) and establishes, for a more balanced comparison, new intercomparable threshold levels for the acceptance of QSAR models as externally predictive. In a precautionary approach, Q^2_{Fn} and $\overline{r^2_m}$ should be raised to 0.70 and 0.65, respectively, instead of the commonly used values (Q^2_{Fn} = 0.6 and $\overline{r^2_m}$ = 0.5), while our preliminary proposal of the CCC threshold of 0.85 is verified to be reasonable by this new approach and, thus, confirmed here. We determined these new intercomparable cutoff values specifying an acceptable level of noise (a measure of the precision) using a simulation protocol that generates a huge number of data sets and not just any QSAR model specifically, which is not necessary for this kind of data agreement evaluation. During this study, we highlighted that the contemporaneous analysis of the scatter plots of experimental vs predicted data is fundamental to accept a model as predictive. Plot methods should be more widely used in model studies, as exemplified by the discovery of some poor scatter plots that show quite good statistical values for some criteria. A scatter plot cannot be a substitute for validation criteria, but it should always accompany their values when checking the predictivity of QSAR models.

From our simulation results, it is evident that only CCC and Q^2_{F3} are always symmetrical with respect to the zero shift/angle applied, reflecting very well also the RMSE symmetry whatever

dx.doi.org/10.1021/ci300084j | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

the shift type (location or/plus scale shifts). This relieves the QSAR developer of being concerned about the sign of the data shift; on the contrary, for the remaining validation criteria, the situation is different for different types of bias. An advantage of CCC in comparison to Q^2_{F3} is that for CCC calculation there is no need for training set information, which is not always available in QSAR. Moreover, CCC can also be used for internal validation to verify model robustness.

Under the scale shift scenario, Q^2_{F1} and Q^2_{F2} have strongly different values for the same RMSE, depending on the side of the bias. Note that $\overline{r^2_m}$ works better in this biased scenario and, in this case, is the most restrictive of the studied validation criteria, while CCC and Q^2_{F3} show a somewhat low sensitivity. In the other two bias types (location and location plus scale shifts), CCC and Q^2_{Fn} perform better in comparison to $\overline{r^2_m}$. Indeed, $\overline{r^2_m}$ was shown to be not sensitive enough to the change in applied bias (consequently, with regression line slopes k and k' not near 1), thus with a tendency to accept not predictive models, meaning models that could be rejected looking at the corresponding graphs and are effectively rejected by the other more restrictive criteria.

Our study, through many different simulations and not just on a particular QSAR model, leads us to advise caution when considering the reliability of $\overline{r^2_m}$ in cases of location and location plus scale shifts and of $Q^2_{F1/2}$ in cases of scale shift, meaning shifts which are not *a priori* known but could be evident just by looking at the corresponding QSAR model plots.

The strongest biases in our simulated data are, hopefully, not usual in QSAR studies. They are extreme cases, but nevertheless, each validation criterion must guarantee its reliability and sensitivity in all scenarios. Also, its behavior must guarantee to have no drawbacks or pitfalls, not even in extreme situations.

In conclusion, given the different behaviors of the various validation criteria in various potential cases of bias and because it is not possible to know a priori the bias, it is not easy to select the "best" external validation criterion for QSAR models before having looked at the model scatter plots. Thus, even if we have now demonstrated that CCC and Q^2_{F3} are the more reliable in all the studied situations, we suggest the use of more than a single criterion for the acceptance of QSAR models as externally predictive (regardless of the number, size, and composition of the external data sets) and to contemporaneously always visualize the corresponding scatter plots. Our in-house software for QSAR model development and validation, QSARINS[24] (soon freely available for academia), implements both the calculation of all the validation criteria and the corresponding plots.

■ ASSOCIATED CONTENT

Supporting Information

Supporting Information SI1: Explanation SI1-1 and Figure SI1-1 for the introductory discussion on the slopes (k and k'), plus additional explanations (SI1-2 to SI1-6) and Figures SI1-2 to SI1-4 on the study of the validation criteria. Supporting Information SI2: Figures SI2-1 to SI2-3: Plots of case study models and relevance of regression line slopes k and k'. Supporting Information SI3: Additional notes on the case study. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*Tel: +39-0332-421573. Fax: +39-0332-421554. E-mail: paola.gramatica@uninsubria.it. Web site: http://www.qsar.it.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors thank Dr. Stefano Cassani for his collaboration in verifying the Roy calculations and Prof. Giorgio Binelli and Ester Papa for helpful discussions. This work was supported by the European Union through the project CADASTER FP7-ENV-2007-1-212668.

■ REFERENCES

(1) Golbraikh, A.; Tropsha, A. Beware of q^2. *J. Mol. Graphics Modell.* **2002**, *20*, 269−276.

(2) Tropsha, A.; Gramatica, P.; Gombar, V. K. The importance of being earnest: Validation in the absolute essential for successful application and interpretation of QSPR models. *QSAR Comb. Sci.* **2003**, *22*, 69−76.

(3) Gramatica, P. Principles of QSAR models validation: Internal and external. *QSAR Comb. Sci.* **2007**, *5*, 694−701.

(4) Tropsha, A. Best practices for QSAR model development, validation, and exploitation. *Mol. Inf.* **2010**, *29*, 476−488.

(5) Aptula, O. A.; Jeliazkova, N. G.; Schultz, T. W.; Cronin, M. T. D. The better predictive model: q^2 for the training set or low root mean square error of prediction for the test set? *QSAR Comb. Sci.* **2005**, *24*, 385−396.

(6) Puzyn, T.; Gajewicz, A.; Rybacka, A.; Haranczyk, M. Global versus local QSPR models for persistent organic pollutants: Balancing between predictivity and economy. *Struct. Chem.* **2011**, *22*, 873−884.

(7) Puzyn, T.; Mostrag-Szlichtyng, A.; Gajewicz, A.; Skrzyński, M.; Worth, A. P. Investigating the influence of data splitting on the predictive ability of QSAR/QSPR models. *Struct. Chem.* **2011**, *22*, 795−804.

(8) Oberg, T. A QSAR for the hydroxyl radical reaction rate constant: Validation, domain of application, and prediction. *Atmos. Environ.* **2005**, *39*, 2189−2200.

(9) Shi, L. M.; Fang, H.; Tong, W.; Wu, J.; Perkins, R.; Blair, R. M.; Branham, W. S.; Dial, S. L.; Moland, C. L.; Sheehan, D. M. QSAR models using a large diverse set of estrogens. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 186−195.

(10) Schüürmann, G.; Ebert, R.; Chen, J.; Wang, B.; Kühne, R. External validation and prediction employing the predictive squared correlation coefficients test set activity mean vs training set activity mean. *J. Chem. Inf. Model.* **2008**, *48*, 2140−2145.

(11) Consonni, V.; Ballabio, D.; Todeschini, R. Comments on the definition of the Q^2 parameter for QSAR validation. *J. Chem. Inf. Model.* **2009**, *49*, 1669−1678.

(12) Consonni, V.; Ballabio, D.; Todeschini, R. Evaluation of model predictive ability by external validation techniques. *J. Chemom.* **2010**, *24*, 194−201.

(13) Roy, K. On some aspects of validation of predictive quantitative structure−activity relationship models. *Expert Opin. Drug Discovery* **2007**, *2*, 1567−1577.

(14) Ojha, P. K.; Mitra, I.; Das, R. N.; Roy, K. Further exploring r^2_m metrics for validation of QSPR models. *Chemom. Intell. Lab. Syst.* **2011**, *107*, 194−205.

(15) Ojia, P. K.; Roy, K. Comparative QSARs for antimalarial endochins: Importance of descriptor-thinning and noise reduction prior to feature selection. *Chemom. Intell. Lab. Syst* **2011**, *109*, 146−161.

(16) Chirico, N.; Gramatica, P. Real external predictivity of QSAR models: How to evaluate it? Comparison of different validation criteria and proposal of using the concordance correlation coefficient. *J. Chem. Inf. Model.* **2011**, *51*, 2320−2335.

(17) Lin, L. I. A concordance correlation coefficient to evaluate reproducibility. *Biometrics* **1989**, *45*, 255−268.
(18) Roy, K.; Mitra, I.; Kar, S.; Ojha, P.; Das, R. N.; Kabir, H. Comparative studies on some metrics for external validation of QSPR models. *J. Chem. Inf. Model.* **2012**, *52*, 396−408.
(19) Pogliani, L.; de Julián-Ortiz, J. V. Plot methods in quantitative structure−property studies. *Chem. Phys. Lett.* **2004**, *393*, 327−330.
(20) Emili Besalú, E.; de Julián-Ortiz, J. V.; Pogliani, L. Trends and plot methods in MLR studies. *J. Chem. Inf. Model.* **2007**, *47*, 751−760.
(21) Atkinson, G.; Nevill, A. Comments on the use of concordance correlation to assess the agreement between two variables. *Biometrics* **1997**, *53*, 775−777.
(22) Lin, L. I.; Chinchilli, V. Rejoinder to the letter to the editor from Atkinson and Nevill. *Biometrics* **1997**, *53*, 777−778.
(23) Lei, B.; Yimeng, Ma, Y.; Li, Y.; Liu, H. X.; Yao, X.; Gramatica, P. Prediction of the adsorption capability onto activated carbon of a large data set of chemicals by local lazy regression method. *Atmos. Environ.* **2010**, *44*, 2954−2960.
(24) Chirico, N.; Papa, E.; Kovarich, S.; Cassani, S.; Gramatica, P. *QSARINS, Software for QSAR MLR Model Development and Validation*; University of Insubria: Varese, Italy, 2012. http://www.qsar.it (accessed May 12, 2012).