JOURNAL OF
CHEMICAL INFORMATION
AND MODELING

ARTICLE

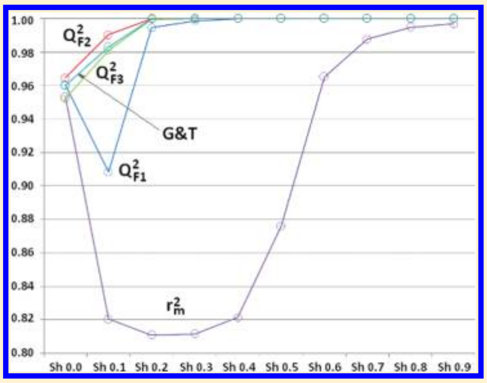
pubs.acs.org/jcim

Real External Predictivity of QSAR Models: How To Evaluate It? Comparison of Different Validation Criteria and Proposal of Using the Concordance Correlation Coefficient

Nicola Chirico[†] and Paola Gramatica[*,†]

[†]QSAR Research Group in Environmental Chemistry and Ecotoxicology, Department of Structural and Functional Biology, University of Insubria, Via Dunant 3, 21100, Varese, Italy

Ⓢ *Supporting Information*



ABSTRACT: The main utility of QSAR models is their ability to predict activities/properties for new chemicals, and this external prediction ability is evaluated by means of various validation criteria. As a measure for such evaluation the OECD guidelines have proposed the predictive squared correlation coefficient Q^2_{F1} (Shi et al.). However, other validation criteria have been proposed by other authors: the Golbraikh-Tropsha method, r^2_m (Roy), Q^2_{F2} (Schüürmann et al.), Q^2_{F3} (Consonni et al.). In QSAR studies these measures are usually in accordance, though this is not always the case, thus doubts can arise when contradictory results are obtained. It is likely that none of the aforementioned criteria is the best in every situation, so a comparative study using simulated data sets is proposed here, using threshold values suggested by the proponents or those widely used in QSAR modeling. In addition, a different and simple external validation measure, the concordance correlation coefficient (CCC), is proposed and compared with other criteria. Huge data sets were used to study the general behavior of validation measures, and the concordance correlation coefficient was shown to be the most restrictive. On using simulated data sets of a more realistic size, it was found that CCC was broadly in agreement, about 96% of the time, with other validation measures in accepting models as predictive, and in almost all the examples it was the most precautionary. The proposed concordance correlation coefficient also works well on real data sets, where it seems to be more stable, and helps in making decisions when the validation measures are in conflict. Since it is conceptually simple, and given its stability and restrictiveness, we propose the concordance correlation coefficient as a complementary, or alternative, more prudent measure of a QSAR model to be externally predictive.

INTRODUCTION

QSAR model validation is fundamental to ensure the reliability of predicted data when applying them to new chemicals. Indeed, to define a model as predictive, it has been demonstrated[1−9] that it is essential to verify the real prediction ability of QSAR models on chemicals never used in model development. This externally validated approach has now been applied in hundreds of more recent QSAR studies.

A QSAR model is calculated using an experimental data set that consists of a number of molecules represented by molecular descriptors (X) and the corresponding responses (Y). During model calculation, as the model is being built up, the key points to verify are model fitting and model robustness, verified respectively by the coefficient of multiple determination R^2 (1), and by internal validation techniques.

The R^2 equation is as follows

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = 1 - \frac{\text{RSS}}{\text{TSS}} \qquad (1)$$

where y_i is the observed dependent variable (the experimental response), $\hat{y}_i$ is the calculated value, $\bar{y}$ is the mean value of the dependent variable, RSS is the residual sum of squares, and TSS is the total sum of squares for n elements of the modeled data set.

Internal validation techniques are based on iterative data set splitting, i.e. a certain number of molecules are involved in the model calculation (training set), while others (called by different names in the literature: test/validation/evaluation set, called here "test set") are in turn put aside and used to check the model's ability to predict them. Cross-validation by the leave-one-out (LOO) method is the most known and used method for calculating internal validation. It evaluates the ability of the model to predict chemicals in the data set one by one, putting them iteratively in the test set.

The formula of the calculated parameter Q^2_{LOO}, widely accepted and applied by the majority of QSAR

Received: May 12, 2011
Published: July 29, 2011

modelers, is

$$Q^2_{\mathrm{LOO}} = 1 - \frac{\sum_{i=1}^{n} (\hat{y}_{i/i} - y_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = 1 - \frac{\mathrm{PRESS}}{\mathrm{TSS}} \qquad (2)$$

where y_i is the observed dependent variable (the experimental response), $\hat{y}_{i/i}$ is the predicted value of the response calculated excluding the i^{th} element from the model computation (excluding more than one element in each iteration is the leave-many-out (LMO) technique), $\bar{y}$ is the mean value of the dependent variable, PRESS is the predictive error sum of squares, and TSS is the total sum of squares for n elements of the complete data set.

Whereas in model development the value of Q^2_{LOO} increases as the number of used molecular descriptors increases (though only to a certain extent as it can then decrease) the fitting measure R^2 continues increasing, and after a certain number of descriptors is reached an overfitted, but not predictive, model can result.

Iterative cross-validations such as LOO and LMO are highly useful to evaluate whether overfitting occurs, and whether the model is robust and stable; however, there is still no agreement among researchers with regard to the reliability such validations have in revealing to what degree the model is able to predict new chemicals completely foreign to the model development process. Thus the real predictivity of a QSAR model must be tested by verifying its performance on completely external chemicals (previously defined[5] and called here 'prediction set', while it is more widely called test set, but in the latest case it can be confused with the test set of the internal validation); nevertheless, some authors consider cross-validation, if properly done, to be sufficient.[10−13]

Predictivity is perhaps the most difficult concept to apply. From a philosophical standpoint, it can be argued that it is impossible to determine an absolute measure of predictivity, as it is greatly dependent on the choice of statistical methods and data sets. Nevertheless, external validation, when performed judiciously, is generally regarded as the most rigorous assessment of predictivity, since predictions are made for chemicals not used in the model development.

There is a fundamental difference between internal and external validation techniques.[3,6] Internal validation, as widely applied, reiteratively excludes one or more molecules (test set), but eventually during the process all the molecules are included in the overall test process (if a compound is in the test set in at least one validation run, no chemical remains "new" at the end of the process). Since all the available information is used, the results regarding model predictivity could be overoptimistic.

Contrary to internal validation, external validation processes never use excluded molecules (prediction set) during model development, which results in a true evaluation of model performance on new chemicals.

Ideally, such a prediction set would consist of new data taken from external sources and found after building the model, but in everyday practice experimental data are often scarce. This is why available data sets are usually split *a priori* by different methods[14−17] even though this leads to the use of reduced subdata sets. This splitting leads to a training set from which the structure−activity relationship is found, and a prediction set of "supposedly unknown" data on which the ability of the developed QSAR model is verified. Again, it is important to remember that the prediction set is completely uninfluential on the presence or absence of a molecular descriptor in the final model, as it never participates in the variable selection process.

The crucial question is as follows: how to verify the real ability of developed QSAR models to predict external data reliably.

In fact, whereas Q^2_{LOO} is a well established and generally recognized measure for internal validation, the situation of external validation measures is not so well-defined and unified.

To verify predictive performance, different validation criteria have been proposed in recent years,[1,7,18−20] and all the proponents have tried to demonstrate the validity and superiority of their proposals.[19−25] Some are based on variation of the Q^2 form,[18−20] while others compare the experimental values of chemicals in the prediction set against corresponding values predicted by the model.[1,7]

State of the Art. It is useful to summarize here the more relevant (and most used) criteria proposed to date for the evaluation of predictive performance on external sets.

The first external validation measure proposed,[4,18] and one widely used by different authors in QSAR literature (also by the author's group, as it was implemented in the software used for QSAR modeling: MOBY DIGS of Todeschini et al.)[26] and also suggested in the OECD guidance document on the validation of QSAR models,[27] has the following formulation

$$Q^2_{\mathrm{F1}} = 1 - \frac{\sum_{i=1}^{n_{\mathrm{EXT}}} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n_{\mathrm{EXT}}} (y_i - \bar{y}_{\mathrm{TR}})^2} \qquad (3)$$

The form is similar to Q^2_{LOO}, but here the sums are over the external prediction set elements and, instead of PRESS (which is calculated using the training set in cross validation), the sum of the squared differences between the prediction set experimental values (y_i) and those calculated by the model ($\hat{y}_i$) is used. In the denominator, the sum of the squared differences of the prediction set experimental values (y_i) and the average of the training set ($\bar{y}_{\mathrm{TR}}$) is used, instead of the plain TSS (which is calculated using the training set values). The use of the average of the training set values, instead of that of the prediction set, is a way of keeping track of the "distance" between the two sets.

Golbraikh and Tropsha[1] proposed a different method to evaluate the differences between the prediction set experimental values and the model predictions. The concept is based on the simple idea that the more the predicted values match the experimental ones, the better the model performance in prediction. Regarding the experimental data points vs those calculated by the model, the determination coefficient in the prediction set (R^2_{ext}) is a straightforward measure of data matching. In fact, the more the determination coefficient is similar to 1, the more the data points lie on the least-squares line. This seems to be reasonable, and a lot of QSAR modelers use this measure to test the predictivity of their models. However, it is important to note that this criterion is not a sufficient measure of agreement, as a perfect match consists only of data lying on a slope 1 line passing through the origin, not just on any slope 1 line but with a different intercept, or any line passing through the origin with a slope different from 1. To solve this, Golbraikh and Tropsha[1] proposed that, in addition to R^2_{ext} (simply called R^2 by the authors), consideration must also be given to the determination coefficient calculated over the external data by forcing the regression line to pass through the origin (this coefficient has been called R_0^2).

Thus it is necessary to be sure that i) the slopes of the regression lines (those related to R^2 and R_0^2) are not too different from 1 and ii) that R^2 and R_0^2 are close enough (closeness is calculated as: $(R^2 - R_0^2)/R^2$). After this calculation, provided that Q^2_{LOO} and R^2 are acceptable, it can be concluded that the experimental and predicted data have a good matching, thus the model can be judged as externally predictive.

It can be noted that in this approach a number of values must be calculated for both axes dispositions (experimental values vs predicted ones and predicted values vs experimental ones), and the authors suggest that the one giving the best result should be considered. Though certainly interesting and rigorous, this method is a little bit complex as more than a single value must be computed, as opposed to Q^2_{F1}.

After the Q^2_{F1} formulation and the proposal of the Golbraikh and Tropsha method, Schüürmann et al. proposed an alternative criterion:[19]

$$Q^2_{\mathrm{F2}} = 1 - \frac{\sum_{i=1}^{n_{\mathrm{EXT}}} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n_{\mathrm{EXT}}} (y_i - \bar{y}_{\mathrm{EXT}})^2} \tag{4}$$

It differs from Q^2_{F1} because the average value at the denominator is calculated using the prediction data set instead of the training one.

The method takes no account of the "distance" from the average of the training values, gaining independence from it. Opinions differ[19−21] as to whether or not this is an advantage, but, in the event that the original training set used to build the model is not available to the user, Q^2_{F2} is advantageous. Another point, highlighted by Schüürmann[19] concerns the fact that $Q^2_{\mathrm{F2}} \leq Q^2_{\mathrm{F1}}$, thus Q^2_{F1} usually gives more optimistic judgements, and, as a consequence, the risk that models not able to make good predictions are accepted is higher. In addition Q^2_{F1}, even if rare, could be even greater than R^2, leading to the contrasting conclusion that the model is able to predict new data better than fitting available ones.

In the same period, the Roy group proposed,[7,22] and applied in two different versions,[23,24] a new external validation metric very similar to the Golbraikh and Tropsha concept but simpler as the result consists of only one value. The following is the last revised version:[23,24]

$$r^2_{\mathrm{m}} = r^2(1 - \sqrt{r^2 - r_0^2}) \tag{5}$$

where r^2 and r_0^2 correspond to R^2 and R_0^2 in the Tropsha and Golbraikh method. This formula can be applied for both external and internal validation: here we concentrate on the external validation form. As can be noted, the distance between the R^2 and R_0^2 determination coefficients is the key in the formula. However, it is important to highlight that in the calculation of r^2_{m} the Roy group does not take into account the R^2 and R_0^2 slopes of the corresponding regression lines, so it is not clear what the reliability of the r^2_{m} value could be when R^2 and R_0^2 give good results but the data points do not match in the predicted/experimental graph (for example, a linear relationship passing through the origin, where the data points on the ordinate values are ten times the ones on the abscissa, leads to a perfect $r^2_{\mathrm{m}} = 1$, because also R^2 and R_0^2 equal 1). In addition, while the R^2 values match regardless the axes disposition used to calculate them (calculation of R^2 using the experimental values vs the predicted ones or, alternatively, the predicted values vs the experimental ones), the R_0^2 values are usually different in the two axes dispositions: this topic was considered only very recently by the Roy group.[25] Even if in this latest paper it is now correctly stated that a specific choice of one axes disposition is arbitrary, so both axes dispositions must be considered, the main problem of r^2_{m} remains the fact that the slopes of the regression lines are still not taken into account. In fact, for the example of a linear relationship passing through the origin with a perfect $r^2_{\mathrm{m}} = 1$ (that reported above in parentheses), also the new proposal[25] to use the average and the difference between the r^2_{m} values for each axes disposition is not useful, as the average will be 1 and the difference will be 0, thus evaluating, as perfectly externally predictive, a model where the predicted data do not completely match the experimental data. This observation raises the problem of the acceptance of some QSAR models externally validated by a single validation criterion with some drawbacks: if about 100 QSAR models in the literature, cited in ref 25, have been validated using only r^2_{m}, some of them could be less predictive than stated or even not predictive .

In the present work, taking into account the above problems related to r^2_{m}, we followed as a guideline the Golbraikh and Tropsha method,[1] as these authors have treated this topic extensively, suggesting the evaluation for predictivity be based on choosing the axes disposition that gives the best results. Thus, the external validation metric r^2_{m} was calculated using both axes dispositions, and then the one giving the best value was compared with the other criteria for model acceptance or rejection.

After the first r^2_{m} formulation,[7,22] Consonni et al. proposed[20,21] a new external validation measure Q^2_{F3}, comparing and highlighting differences with Q^2_{F2}. One preliminary comment regards the relationship $Q^2_{\mathrm{F2}} \leq Q^2_{\mathrm{F1}}$, discussed by Schüürmann,[19] which, though true, does not necessarily imply that Q^2_{F2} is a correct way to estimate the ability of a model to predict. In addition, the fact that there is no reference to the model training set (that is $\bar{y}_{\mathrm{TR}}$) is, according to Consonni et al., a drawback, contrary to Schüürmann's opinion. Moreover, in Consonni's analyses, Q^2_{F1} and Q^2_{F2} were revealed to be biased depending on the data distribution.

In order to overcome these drawbacks the Q^2_{F3} validation criterion, alternative to Q^2_{F1} and Q^2_{F2}, was proposed:[20]

$$Q^2_{\mathrm{F3}} = 1 - \frac{[\sum_{i=1}^{n_{\mathrm{EXT}}} (\hat{y}_i - y_i)^2]/n_{\mathrm{EXT}}}{[\sum_{i=1}^{n_{\mathrm{TR}}} (y_i - \bar{y}_{\mathrm{TR}})^2]/n_{\mathrm{TR}}} \tag{6}$$

As can be seen, the formulation differs from both Q^2_{F1} and Q^2_{F2} as the denominator is calculated on the training set, and both numerator and denominator are divided by the number of the corresponding elements.

By simulations they demonstrated that the result obtained by Q^2_{F3} is always the same, independently of the prediction set distribution (i.e., it is invariant to the measured sampling). In addition Q^2_{F3} seems to be independent of sampling size, while the predictive ability measured by Q^2_{F2} increases as the prediction set object number increases (moreover Q^2_{F2} converge to Q^2_{F1} as the prediction set size increases and also Q^2_{F1} and Q^2_{F2} converge to Q^2_{F3}[20] when the prediction set increases up to a certain level).

Assuming that external test objects are independent of each other, Consonni et al.[20] also demonstrated that the Q^2 value, calculated using the entire external data set, and the average of the Q^2 values, obtained taking separately each external test set object one at a time, should coincide (ergodic property); only Q^2_{F3} fulfills this property, considered relevant by the authors. Moreover, the authors pointed out[21] that since Q^2_{F3} includes information concerning the training set in the formula it could not be rigorously considered a perfect external validation measure, even though the predictions are really obtained by external test data.

A very simple measure of the model's ability to predict, widely used by QSAR modelers as an additional validation criterion, and suggested by Aptula et al.[28] as the criterion of choice for checking external predictivity, is the root-mean-square error in prediction, formulated as:

$$\mathrm{RMSEP} = \sqrt{\frac{\sum_{i=1}^{n_{\mathrm{EXT}}} (\hat{y}_i - y_i)^2}{n_{\mathrm{EXT}}}} \tag{7}$$

that measures the discrepancies among the experimental values vs the ones predicted by the model. It is important to note that this criterion is not useful to evaluate and compare different models because it is dependent on the scale of measure of the modeled responses (thus it is not a relative measure). Consonni et al.[20] state that Q^2_{F3} is able to reproduce the RMSEP ranking, providing the training set has been fixed, while the prediction set varies due to the random sampling.

However, even though RMSEP does not suffice for different model comparisons, it must always be calculated for both the training and the prediction sets, as an additional criterion for the absolute measurement of the error, defined in the domain of the measurement unit. The stability of RMSE for training and prediction sets can also be considered a measure of the model's generalizability. A similar criterion for the absolute measurement of error is MAE (Mean Absolute Error):

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n_{\mathrm{EXT}}} |y_i - \hat{y}_i|}{n_{\mathrm{EXT}}} \tag{8}$$

This survey of various validation criteria gives rise to a crucial question that concerns the choice of the one(s) that should be used to test the real external predictivity of QSAR models. In alternative, a comparison of various measures would be more informative, but a strong disagreement among the criteria would result in further uncertainties.

In our QSAR modeling we verified that these measures, with the threshold values defined by the proponent or widely used by the QSAR community, were not always in good agreement, not only as a quantitative measure of predictivity, but also sometimes for the resultant decision with regard to the real prediction ability of the developed models (acceptance of models as really predictive). We found some models predictive according to one measure, but they had to be discarded as they were not acceptably predictive for other criteria. Thus, due to this verified uncertainty, we recently externally validated our models by contemporaneously using more than one measure, and we have proposed as valid only those models that are considered predictive by all the used criteria.[17,29−33]

Proposal of a Simpler Criterion for External Validation. From the previous analysis it is evident that two types of validation criteria have been proposed: those based on variations of the Q^2 form[18−21] and those based on the difference between the experimental and the predicted data (by the model) of the prediction set.[1,7,22−25] In our opinion the second technique is more intuitive and conceptually simpler and should suffice to evaluate the model performance in terms of relative measurement of errors, so we will concentrate on this option. It is important to remember that an absolute measure of errors, such as RMSE, is always needed for a complete assessment of the model predictivity.

The present paper compares the published and most used external validation measures using, where available, the threshold values defined by each proponent and those widely applied by QSAR modelers in a big simulation exercise to verify whether a simpler (but still reliable) measure can be proposed as an additional measure or even as a substitute for all the others. A basic assumption for a simple check of model predictivity is that the data predicted by any QSAR model for external chemicals, independently of the training data used to develop the model, should, at the maximum level, be in full agreement with the real data of those chemicals, which were never used in the model development. Given this assumption, another simple measure that could be considered is the Pearson's correlation coefficient, due to its independence of the measure units, but it cannot be used for the same reasons previously expressed for the determination coefficient (see comments on the Golbraikh and Tropsha method above). Nevertheless a measure similar to the Pearson's correlation coefficient, but one also able to compare data with a slope 1 line passing through the origin, would be attractive because of its simplicity.

Our choice fell on the concordance correlation coefficient (CCC) proposed by Lin,[34,35] here slightly rearranged with respect to the original for easier readability (see the Supporting Information), as it is well suited to measure the agreement between experimental and predicted data, which should be the real aim of any predictive QSAR models:

$$\hat{\rho}_c = \frac{2\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2 + \sum_{i=1}^{n} (y_i - \bar{y})^2 + n(\bar{x} - \bar{y})^2} \tag{9}$$

where x and y correspond to the abscissa and ordinate values of the graph plotting the prediction experimental data values vs the ones calculated using the model (or the opposite disposition: as can be noted from the formula it does not matter), n is the number of chemicals, and $\bar{x}$ and $\bar{y}$ correspond to the averages of abscissa and ordinate values, respectively.

This coefficient measures both precision (how far the observations are from the fitting line) and accuracy (how far the regression line deviates from the slope 1 line passing through the origin, i.e. the concordance line), consequently any divergence of the regression line from the concordance line gives as a result a value of CCC smaller than 1. The key point is that this result is obtained even if the Pearson's correlation coefficient equals 1, i.e. the data perfectly matches any linear relationship;[34] however, in this last case the model can be precise but not accurate.

The first advantage, compared to the more similar Golbraikh and Tropsha method, is its greater simplicity and its independence of the axes disposition. The second is that no training set information is involved, so it can be considered a true external validation measure, independent of the sampled chemical space.

Thus we wanted to verify whether our choice, the concordance correlation coefficient (CCC), could be used, efficiently and very simply, for the external validation of QSAR models. With this verification in mind, this preliminary paper presents a comparative analysis of the above cited external validation criteria, using the threshold values widely used in the literature, and CCC. The analysis uses simulated data sets and real data sets to cover a wide range of situations and to verify the agreement or disagreement of the various validation criteria and their relative level of optimism in such situations.

In addition, as it can be expected that the number of objects in the data sets will play a big role in model and prediction stability, we wanted to verify and compare the stability of the models and all the studied external validation criteria at different data set sizes and prediction set proportions.

Let us summarize our goals:

1) to propose a single statistical validation measure based on a simple concept;
2) to compare the proposed validation measure with all the previously proposed criteria using the usually applied threshold values, verifying their respective optimism or restrictivity for external prediction performance in the present state of the art in the QSAR community;
3) to evaluate, by simulations, model stability on different data set sizes and prediction set proportions;
4) to evaluate the proposed external validation measure performances on the same data sets of point 3;
5) to extract some relevant examples from simulated and real data sets to study possible conflicting results on the studied criteria;
6) to highlight possible problems in some of the used criteria compared here.

METHODS

Generating Simulated Data Sets. The aim is to generate a certain number of simulated data sets of different sizes and different training vs prediction set proportions. Simulated data sets sizes are as follows: 24, 48, 96, 192, 384, 768, and 1536 elements. Prediction set proportions are as follows: 1/2, 1/4, and 1/8.

Simulated data sets are built using one descriptor and the corresponding response. The descriptor values are chosen at random, following a Gaussian distribution, from a range spanning from 0 to 1 as it will soon be explained. Here we recall the Gaussian distribution function in the normalized form:

$$f_{X,\sigma} = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-X)^2/2\sigma^2} \qquad (10)$$

In this work we chose a standard deviation (σ) of 0.15 and an X value of 0.5, thus obtaining a Gaussian distribution centered on 0.5, where the maximum value of the bell shaped distribution lies. We were interested in using the distribution area derived from at least 3σ, that covers 99.7% of the area. The value of 0.15 is arbitrary and chosen just because it rounds $0.5/3 = 0.1\overline{6}$ with the consequence of covering more than the requested 99.7% of the area too.

In order to generate the x values according to the Gaussian distribution we first split the x range (0−1) by the size of each considered data set. Starting from 0 and increasing by the calculated step (1/data set size) up to 1, we extracted at each step a random number from 0 to $\frac{1}{\sigma\sqrt{2\pi}}$. If the random value is smaller or equal to the Gaussian curve value (ordinate), the corresponding x value (abscissa) is considered, otherwise it is rejected and the procedure goes to the next step. It can also be noted that if there is no interest in the distribution area, the procedure can be simplified: the random numbers are extracted from 0 to 1 and only the exponential part of formula 10, i.e. $e^{-(x-X)^2/2\sigma^2}$, need to be calculated. The whole procedure is repeated until the number of required elements, i.e. the data set size, is reached.

The underlying model for the data set generation is *response* = (*descriptor* + *scattering*) + *shift*. The descriptor values are selected by the aforementioned procedure, while scattering and shift are generated as described below.

In the first round of data set generation both scattering and shift equals zero: in this case the results are also used to check if the procedure is working correctly, e.g. all the validation criteria report that the data to be validated perfectly match. Going further with the simulation, the subsequent data sets are generated adding a certain amount of scattering to the responses and then extracting at random the prediction set. The remaining data are used as the training set. After exploring the range of data scattering of interest, new data sets are generated following the same procedure but adding a systematic shift to the prediction set responses. In more details, to generate a single data set a maximum absolute value of scattering is chosen from a range that spans between 0 and 1 using discrete steps of 0.01. Thus, no scattering is added to the first data set, scattering values spanning at random from −0.01 to +0.01 are added to the second one, values from −0.02 to +0.02 to the third one and so on, up to −1 to +1. After doing all the analyses, a new group of data sets is generated, adding a systematic shift value to the prediction set responses. The systematic shifts span from 0 to 0.9 in discrete steps of 0.1, so ten groups are generated. Finally, to mitigate chance results, ten rounds of the whole aforementioned procedure are executed, thus obtaining a total of 210,000 data sets.

For every generated data set an OLS regression model is calculated, using the training set, in order to obtain the model equation (intercept + variable coefficient) to be applied to the prediction set. There is only one variable coefficient because we use only one descriptor; the calculated coefficient is multiplied by the external prediction set descriptors, in order to calculate the external data set predictions.

Validation Criteria Thresholds. In all simulations the validation measures are calculated only if R^2 is greater than 0.7, Q^2_{LOO} is greater than 0.6, and the difference between them is smaller than 0.1. We raised the threshold of the above criteria with respect to the originally suggested values (0.6 and 0.5, respectively) to values commonly applied by the author's group. If these conditions are satisfied (thus the models are robust), the following thresholds are used to determine if the model is able to make good external predictions:

Q^2_{F1}, Q^2_{F2}, and Q^2_{F3} greater than 0.6 (a value in accordance with Q^2_{LOO}), r^2_{m} greater than 0.5 (as suggested by Roy[7]), Golbraikh and Tropsha method (called "G&T" in this work) as suggested in their paper.[1]

An arbitrary cutoff value of 0.85 for CCC proposed here is chosen on the basis of an acceptable level of data scattering of, at most, 0.15. It is important to note that such a CCC value (0.85) is not very restrictive compared to the literature.[34] The commonly used thresholds of each validation measure are applied to all simulated data sets, and then, for every validation criteria, a counting of the accepted and rejected models is performed. Such counting is then compared to verify which one of the aforementioned criteria is the most restrictive.

General Trend of the Validation Criteria. In addition, to study the general behavior of the validation measures as the scattering/shift increases, simulated data sets composed of 10^6 data points are generated for every scattering range value and systematic shift. From this data set 1/10 of the total objects is randomly extracted to build the prediction set. In this case, the model robustness by Q^2_{LOO} is not checked because of the data set size.

Validation Criteria Performance. To evaluate validation measure performance, relatively stable reference values are calculated, averaging the values of the validation criteria of 100 simulated data sets of 2000 elements. For the evaluation of the external validation measures, 1000 elements are extracted at random as the prediction set. The discrepancy, used to evaluate the criteria performance, is calculated as:

$$\text{discrepancy} = \frac{\sum_i |v_{i_{\mathrm{ref}}} - v_i|}{\sum_i |v_{i_{\mathrm{ref}}}|} \tag{11}$$

where i is the scattering index (0 to 99), $v_{i_{\mathrm{ref}}}$ is the reference measure value, and v_i is the measure value to be compared. Criterion performance is regarded as the discrepancy from the reference, thus the higher this value the lower the performance.

Real Data Sets. The mutagenicity potency in TA100 (without the S9 activation system) for 48 nitro-PAHs was obtained from the Benigni Report for OECD[36] and was modeled using GA-OLS by Gramatica et al.[37]

In the GA-population of models for PAH mutagenicity some models appeared stable and predictive by internal validation measures (Q^2 and Q^2_{boot}) but less predictive (or even unpredictive: $Q^2_{\mathrm{EXT}} = 0$) when applied to external chemicals that were really never presented to the GA during model development. These models have been here verified by all the compared validation criteria.

The boiling point data for 93 perfluorinated chemicals, taken from the SRC-PhysProp database, as reported by ChemID plus[38] and implemented with data from Hendricks,[39] were modeled by Bhhatarai and Gramatica by GA-OLS.[33] Some models in the GA population, that had discordant results regarding their external performance, based on the tested validation measures (Q^2_{F1} and Q^2_{F3}), are here verified by all the compared criteria.

RESULTS AND DISCUSSION

The first aim of this work was to compare external validation criteria in two types of comparisons: level of acceptance region and agreement among them. We wanted to verify the real situation as it is in the QSAR world, so we kept the normally applied threshold values for each criteria in our comparative analysis.

The level of acceptance region means that the analytical forms of the validation parameters should tell by themselves, at least in principle, if one of the validation parameters is the most restrictive. Since stochastic variables are used it is not easy to answer this question by simply looking at the formulas, so simulated data are used. In this case it is essential to reduce randomness as much as possible using a very high number of simulated objects (in this case 10^6, see Methods). Randomness is defined here as the random scattering of the data in the simulated data sets. Data scattering, and possible unevenness in their distribution, are expected to impact external validation measures values at different levels because the analytical forms of the criteria differ.

In the study the simulated data were taken at random, so there is no pattern in the training and prediction data. This is quite different from real QSAR modeling, where data including information on molecular structure and response play a dominant role. Indeed, this study was based on unstructured simulated data, so the main focus was not on the QSAR model itself, it was specifically focused on the validation criteria performance at the different levels of data scattering. Thus the use of one or more descriptors does not influence the simulation, but the use of one descriptor alone allows the possibility of ignoring any correlation and/or additional sources of variation.

To calculate and verify the position of the specific thresholds of the acceptance regions different values of scattering and systematic shifts are added to the responses of the simulated data sets. In a perfect model all the calculated values match the prediction set values perfectly. In such a case the simulated data set has no scattering or systematic shift added to the prediction set. Keeping the systematic shift to zero and increasing the scattering in the responses in both the training and the prediction sets, some validation criteria should differ in performance. Since, as previously indicated, some validation measures depend on the average of both the training and the prediction responses, different systematic shifts are also added to the prediction set responses. Overall, adding both scattering and systematic shifts should lead to differences in criteria values.

Agreement means the verification of how many times the proposed validation measures agree to accept or reject a model. Since in real situations available data, used to build and test a model, are usually relatively few, the distribution of the data between the prediction and training sets is important in determining the external criteria values. In fact, for some of the validation measures the average experimental values in the training and prediction sets play a big role. Sometimes, some of these measures disagree in accepting a model so, by means of simulated models, we quantify how many times all the criteria agree to accept/reject a model, and all the situations in between (some agree and some disagree).

To study the validation measures agreement, a more elaborated and realistic framework is generated. This includes model cross-validation and the generation of data sets of realistic size, both small and relatively big. It can be expected that the smaller data set sizes, compared with those used for the analysis of the acceptance region, the training and prediction set data may not be uniformly distributed, as usually happens in real data sets. In such data sets, especially the smaller ones, the prediction set proportion is important (there is a trade-off between a good fitted and stable model and the capability of prediction, because available data are scarce). For this reason classes of different prediction set proportion for each data set size are generated.

General Validation Criteria Performance. The first step is to make a comparison of the behavior of CCC and the other validation measures, using data sets of 10^6 data points. In this comparative exercise the Q^2 validation criteria are rewritten in the following forms:

$$Q^2_{\mathrm{F1}} = 1 - \frac{\left[\sum_{i=1}^{n_{\mathrm{EXT}}} (\hat{y}_i - y_i)^2\right] / n_{\mathrm{EXT}}}{\left[\sum_{i=1}^{n_{\mathrm{EXT}}} (y_i - \bar{y}_{\mathrm{TR}})^2\right] / n_{\mathrm{EXT}}} \tag{12}$$

$$Q^2_{F2} = 1 - \frac{[\sum_{i=1}^{n_{EXT}} (\hat{y}_i - y_i)^2]/n_{EXT}}{[\sum_{i=1}^{n_{EXT}} (y_i - \bar{y}_{EXT})^2]/n_{EXT}} \tag{13}$$

$$Q^2_{F3} = 1 - \frac{[\sum_{i=1}^{n_{EXT}} (\hat{y}_i - y_i)^2]/n_{EXT}}{[\sum_{i=1}^{n_{TR}} (y_i - \bar{y}_{TR})^2]/n_{TR}} \tag{14}$$

In order to obtain an analytical form similar to Q^2_{F3}, the numerators and the denominators of Q^2_{F1} and Q^2_{F2} are divided by n_{EXT} (note that in this way Q^2_{F1} and Q^2_{F2} are multiplied by one), thus all the numerators are identical. Looking at the denominators, and randomly extracting a large number of data values from the same range, it is expected that $\bar{y}_{TR}$ and $\bar{y}_{EXT}$ will be similar. Indeed, dividing the total sum of the squares at the denominator by the respective number of elements, and as the data range is the same, will result in all the denominators having similar values.

The level of convergence among the validation criteria, demonstrated by the analytical formulas, is checked further with increasing scattering values, including also r^2_m, the Golbraikh and Tropsha method, and CCC for comparison: the results related to simulations without any systematic error in the responses (shift 0.0) are shown in Figure 1, where the normally used cutoff values are also reported to highlight the acceptance/rejection regions, based on usual practices in QSAR modeling.

As expected, also from the formulas and already pointed out by Consonni et al.[20,21] Q^2_{F1},Q^2_{F2}, and Q^2_{F3} match. When r^2_m is calculated using the predicted external data on the abscissa axes and the experimental external data on the ordinate (r^2_m-EyPx), r^2 and r_0^2 match (or nearly so), as a consequence of the simulation protocol, thus zeroing the square root value. In that case r^2_m equals r^2 in eq 5, which results in the same form of Q^2_{F2} (eq 4), as a consequence r^2_m and Q^2_{F2} graphs match. However, all the Q^2_{Fn} have the usual cutoff of 0.6 (circled 2 in Figure 1), while r^2_m-EyPx has cutoff of 0.5, proposed by Roy[7] (circled 3 in Figure 1).

The Golbraikh and Tropsha method is not as easy to graph as the aforementioned criteria because it consists of a number of different values. Indeed, it is only after having calculated all the values that the user is able to determine if a model can be accepted as predictive or not. A rejection region was here identified by our simulation: after a scattering value of 0.2 the Golbraikh and Tropsha method rejected all the models.

A concordance correlation coefficient cutoff value of 0.85 was arbitrarily chosen (circled 1 in Figure 1) considering, in this preliminary paper, that a model with a scattering value greater than 0.15 is not predictive. After plotting all the external validation measure values and reporting the relative cut-offs, normally used in daily QSAR practice (see Figure 1), CCC proved to be the most restrictive in accepting models, followed by the Golbraikh and Tropsha method, then by all the Q^2 validation criteria, and last by r^2_m-EyPx. In this case the combination of axes giving the highest r^2_m value is chosen because the same reasoning as for the Golbraikh and Tropsha method[1] is applied (i.e., the best axes combination). The other axes combination, r^2_m-ExPy, proved to be almost as restrictive (circle 4) as CCC, but remains unjustified as the authors did not explore this topic in their papers[7,22−24] (i.e., why should one axes combination always be chosen instead of the other?).

Figure 1 suggests that the threshold values, commonly used by the proposed validation criteria, should be raised (e.g., all Q^2_{Fn} from 0.6 − the threshold used here - to 0.75, and r^2_m-EyPx from 0.5 to 0.75). Thus, the QSAR models, accepted until now as predictive by applying the proposed cutoff values, could be defined as not predictive by the new threshold values. In any case, the aim of this preliminary work was not to propose new thresholds for the validation criteria, so we concentrated on those

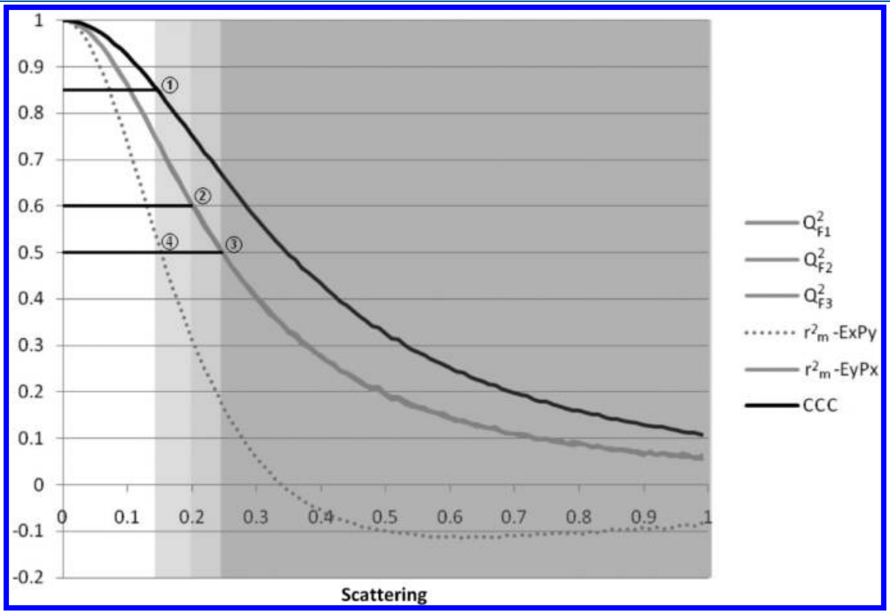


Figure 1. Validation criteria comparison calculated using the simulated data sets (shift 0.0). Commonly applied thresholds are in the encircled numbers: 1) CCC − rejection region starting from light gray, 2) Golbraikh and Tropsha method, Q^2_{F1}, Q^2_{F2}, Q^2_{F3} − rejection region starting from middle gray, 3) r^2_m-EyPx (predicted data on the abscissa and experimental values on the ordinate) − rejection region in dark gray, 4) r^2_m-ExPy (experimental data on the abscissa and predicted values on the ordinate).

more commonly used, for comparison purposes. The determination of new acceptable threshold values for each coefficient will be the topic of a future paper, presently in preparation.

Since it can be expected that the main differences, at least among the Q^2 validation criteria, arise when the training and prediction set averages differ (see the Q^2 formulas), different values of systematic shifts are added to the prediction set responses. In this way, since 10^5 elements are used as the prediction set (see Methods), it is practically ensured that the average value of the training set responses is different from that of the prediction set. Adding a systematic shift of 0.1 (see Supporting Information Figure SI1a) CCC, Q^2_{F2}, Q^2_{F3}, and the Golbraikh and Tropsha method (not reported in the graph) exclude all the models at all levels of scattering, while Q^2_{F1} and (in a bigger amount) r^2_m accept a certain number of them. Q^2_{F1} was already considered to be overoptimistic by Schüürmann et al.,[19] and our analysis supports this. Since the external responses are systematically shifted, Q^2_{F1}, Q^2_{F2}, and Q^2_{F3} no longer match, as expected, especially Q^2_{F1} with respect to Q^2_{F2} and Q^2_{F3} (Q^2_{F2} and Q^2_{F3} graphs are very similar but do not coincide, this is enlarged upon in Figure SI1b). In summary, when a systematic shift of 0.1 is added, only Q^2_{F1} and mainly r^2_m accept models, as they are probably too optimistic. At this systematic shift the remaining external validation techniques are similarly restrictive, as none of the models are accepted.

To test the overoptimistic scenarios posed by r^2_m and Q^2_{F1}, the systematic shift is increased in discrete steps of 0.1. As a general result, Q^2_{F1} does not accept any model starting from a level of systematic shift of 0.2 (see Supporting Information Figure SI1c), while r^2_m does not accept any model at a very higher shift value of 0.6 (see Supporting Information Figure SI1d and SI1e), so it is confirmed that r^2_m is the most optimistic validation measure among the ones studied. Overall, CCC, Q^2_{F2}, Q^2_{F3}, and the Golbraikh and Tropsha method are more restrictive in comparison to Q^2_{F1} and r^2_m; in particular, CCC is the most restrictive, thus more precautionary in accepting a model as externally predictive.

Validation Criteria Comparison Using Random Data Sets of Realistic Sizes. Using the CCC threshold of 0.85, the accordance (in accepting or rejecting a model) between this coefficient and all the others is verified by generating 210,000 data sets of different sizes and different training-prediction set proportions (see Methods).

Plotting the agreement of CCC with the other validation measure vs the systematic shift added to the prediction set responses led to the graph of Figure 2a. Without adding any systematic shift, an agreement proportion value of (around) 0.96 between CCC and the other validation measures is obtained (only Q^2_{F3} and r^2_m are just below this value). As the systematic shift increases, the agreement with r^2_m drops to 0.82 and that with Q^2_{F1} to around 0.91, while the other validation criteria increase their agreement further, reaching up to at least 0.98. Excluding r^2_m, starting from a shift value of 0.2, the agreement of the validation measures with CCC raises to almost 1.00. Focusing on r^2_m, the agreement follows a "U" shaped curve where the values fall below 0.82 in the range of shifts of 0.2−04 and then raise to almost 1.00 at high shift values (0.8 and 0.9, where presumably all the criteria agree in rejecting the models). In Figure 2b the proportion of models accepted as predictive by the different measures over the total generated by the simulation with various shifts is plotted. If no shift is added, the accordance among CCC and the other criteria is reasonably based on accepting or rejecting a model (see left part of Figure 2b, where it can be noted that the concordance coefficient is the most severe). As the shift to the prediction set responses is increased up to 0.1, the percentage of accepted models differs among the validation measures, so this region can be considered more critical. As the shift increases further, also the agreement among the validation measure increases, but it is mainly due to rejecting the models (see right part of Figure 2b). As it can be seen, there is very good agreement among the validation criteria and CCC, except for r^2_m in almost all shifted responses and also Q^2_{F1} for a shift value of 0.1.

If all validation measures agree with CCC, there are basically no uncertainties, but questions arises when one or more criteria are not consistent with each other. To evaluate this topic the number of the measures that accept-reject a model are counted organizing the situations in the following categories: A (4−1), B (3−2), C (2−3), D (1−4), where the first number in parentheses is the number of validation criteria that accept a model, while the second one is the number of validation criteria that reject the model. The proportion of agreement with CCC and the proposed categories is calculated, obtaining the following values: category A = 22%, category B = 14%, and zero for the other two categories. As a consequence it is again verified, in an other way, that CCC is more restrictive when the validation measures are

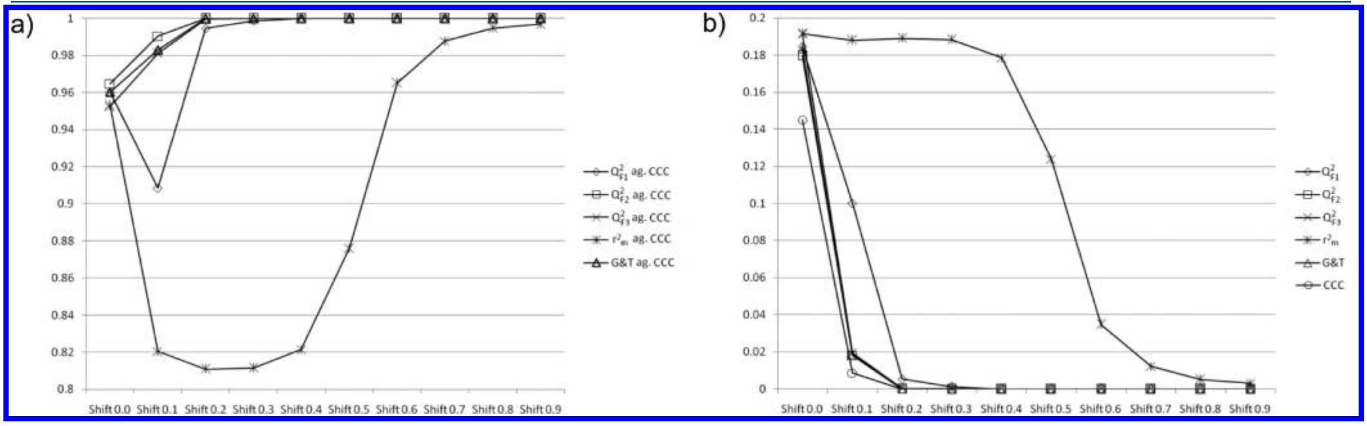


Figure 2. a) Proportion of agreement of CCC with other external validation criteria in accepting or rejecting models, with various level of systematic shifts. b) Proportion of models accepted as predictive by the different criteria over the total generated by the simulation, with various level of systematic shifts.

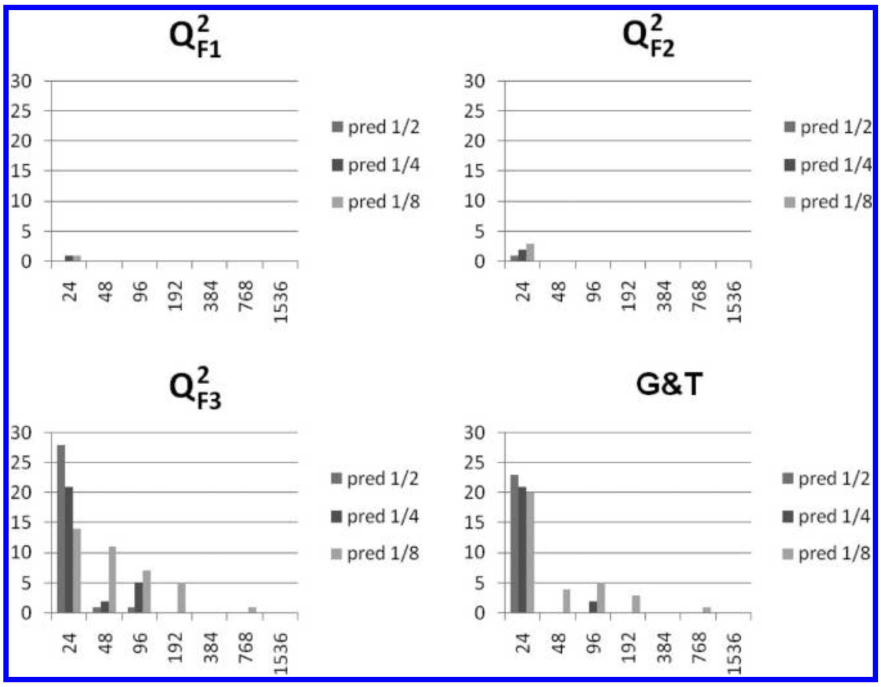


Figure 3. Number of models rejected by at least one of the external validation criteria while CCC accepts them (Case 1). The number of models is on the ordinate axis, while the data set size is on the abscissa axis. Different gray levels represent different prediction set proportions extracted from the simulated data sets.

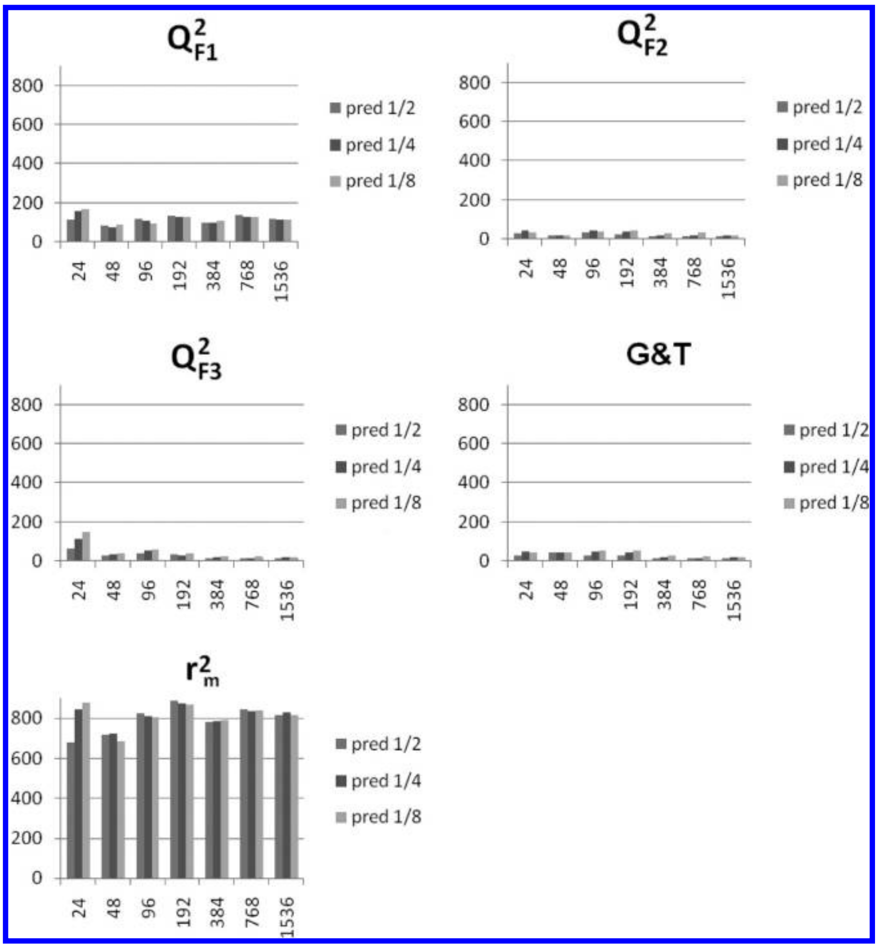


Figure 4. Number of accepted models by at least one of the external validation criteria while CCC reject them. The number of models is on the ordinate axis, while the data set size is on the abscissa axis. Different gray levels represent different prediction set proportions extracted from the simulated data sets.

not consistent in accepting a model. Looking further at category C it can be again confirmed that Q^2_{F1} and r^2_m are probably too optimistic in accepting too many models as predictive. r^2_m is even more permissive in Category D (see Tables SI1 and SI2).

The validation criteria compared against CCC are further explored, splitting the inconsistencies in two cases: 1) CCC accepts the model while at least one of the other validation measures rejects it and 2) CCC rejects the model while at least one of the other validation measures accepts it. This analysis is performed, for every data set size and prediction set proportion, to check if any relevant trend occurs.

Case 1 (Figure 3) shows very few inconsistencies (compared to the total number of 210,000 generated data sets) and, where they are found, there are either too few data points for the calculation of CCC (at least 10 are probably needed[35]) or the data points are peculiarly arranged. Since the number of inconsistencies is very small considering the large amount of models computed in this simulation, the possibility of accepting a "bad" model, evaluating it as externally predictive by CCC, seems to be somewhat remote. Concerning r^2_m, no inconsistencies with CCC are found for case 1. Overall, we have verified that there are rare cases, mainly in small data sets, where CCC accepts a model, while Q^2_{F3} and the Golbraikh and Tropsha method consider it as not predictive.

On the other hand, case 2 reports much more inconsistencies mainly with r^2_m and, in lesser amount, Q^2_{F1} (Figure 4), and it can be noted that there is no improvement in using bigger data sets and/or different prediction set proportions. This observation allows the conclusion that the number of models accepted by the validation criteria is invariant to the prediction set proportion and the data set size, as no relevant trend is evident.

Detailed Comparative Analysis of the External Validation Criteria Stability. A more detailed study of all the validation criteria applied to the above realistically sized data sets is performed with increasing scattering and using different training/prediction set proportions. The average and the standard deviation for each external validation measure is calculated over ten rounds of simulated data sets of size from 24 to 1536 elements (Supporting Information Figures SI2a to SI2g). Looking at the corresponding graphs, the trend of all the validation measures, until the data set size of 48 elements, is very irregular, thus suggesting that data sets of similar sizes should be considered with caution during model development (see Figure 5 upper part). The data sets of 96 elements are probably not sufficiently regular, while data sets of 192 elements (see Figure 5 lower part) seem to be a good compromise between the number of elements and the erratic behavior of the validation measures found in the smaller data sets. As expected, the irregularity of

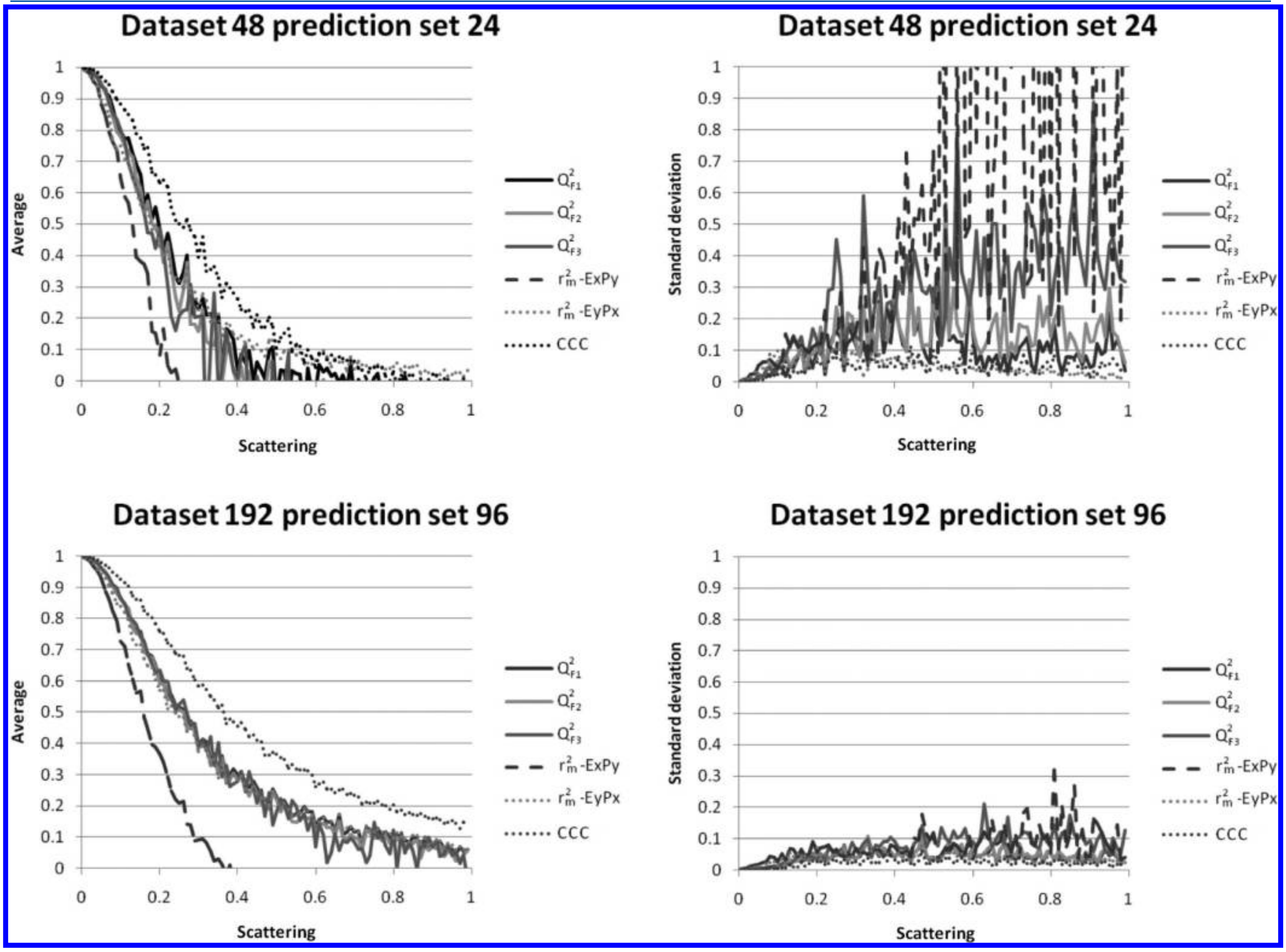


Figure 5. Examples of external validation criteria stability calculated over ten rounds of simulated data sets.

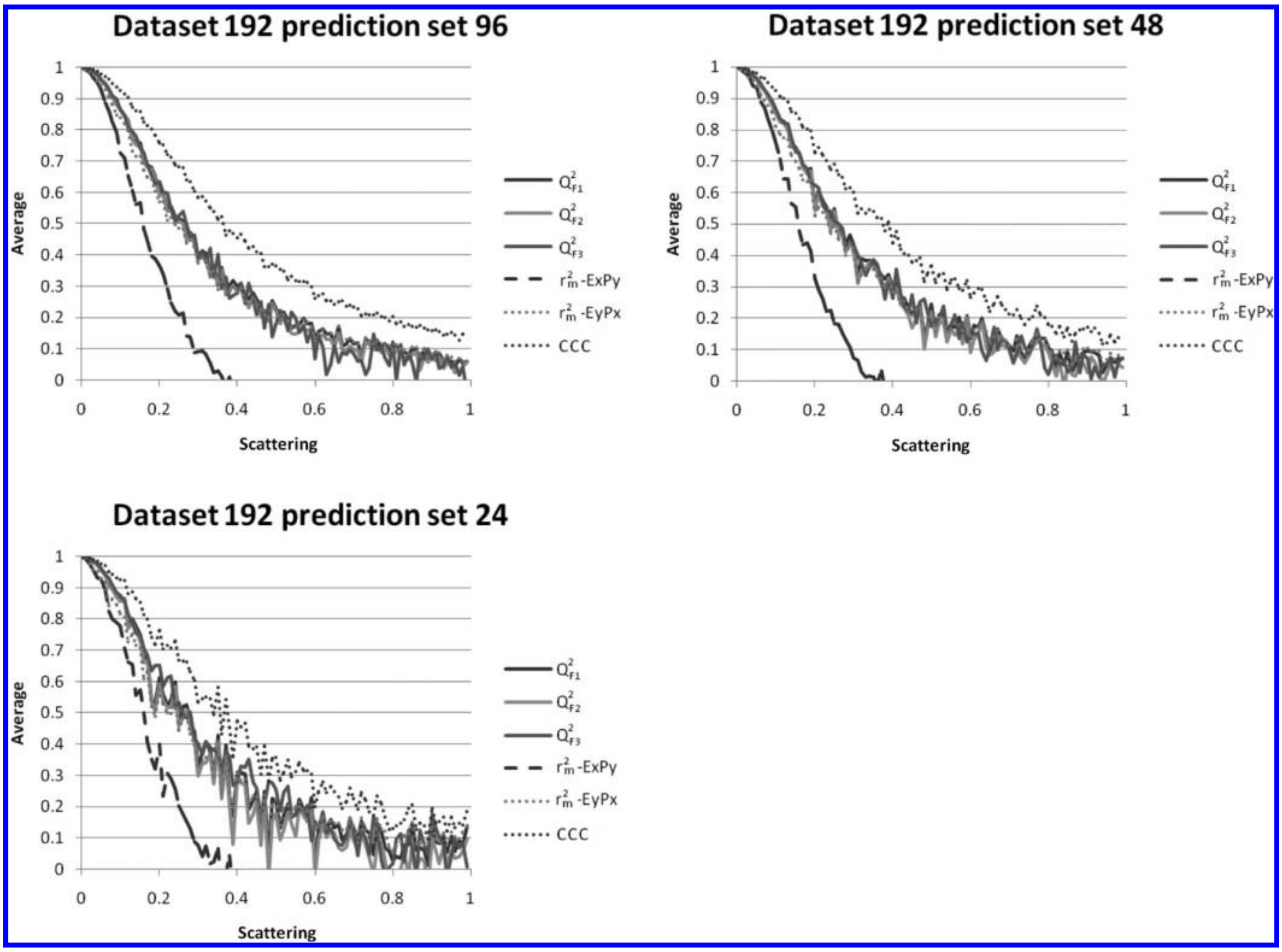


Figure 6. Example of different regularity of the external validation criteria at different prediction set proportions.

the validation criteria decreases as the data set size increases further.

Looking at all the SI2 graphs it can be noted that, for each data set size, starting from the 1/2 prediction set proportion to 1/8, the validation measures curves become more and more irregular (e.g., Figure 6). This result is supported by the statistical theory which demonstrates that the relative standard deviation in the RMSEP estimate decreases as the number of prediction samples increases.[41] The observed trends of the external validation criteria should derive mainly from the external data distribution and the corresponding number of elements (the fewer they are, the more different the external validation criteria should appear at different simulation rounds).

It is evident from this analysis that the size of the prediction set is a dominant element in determining the regularity of the model prediction. However, it is also important to remember that in QSAR modeling the size of the training set is also relevant, because it is expected that the bigger the training set (thus the included information) the better the model's stability and predictive ability.

In order to verify whether the observed irregularities could also be a consequence of model instability, we focused on Q^2_{LOO} averages and standard deviations. Plotting the Q^2_{LOO} values vs the scattering values reveals that the bigger the training set proportion (and the corresponding test set size for the calculation of Q^2_{LOO}) the smaller the derived Q^2_{LOO} irregularity. An example of the result of this analysis is presented in Figure 7, and all the results are shown in Supporting Information Figure SI3a.

Additionally, the performance of the external validation criteria is evaluated by verifying the discrepancy from reference values. The reference values are calculated averaging 100 data set of 2000 elements of which 1000 were used as prediction set (see Supporting Information Figure SI4a and Figure SI4b and Figure 8 as an example).

It can be seen for Q^2_{F1} and Q^2_{F2} that the bigger the prediction set proportion the better the criteria performance, suggesting that these parameters are particularly sensitive to prediction set proportions, smaller prediction sets performing worse. Q^2_{F3} and r^2_m (calculated using the experimental values on the abscissa axis: ExPy) had not a similar trend, instead r^2_m (calculated using the experimental values on the ordinate axis: EyPx) and CCC perform similarly to Q^2_{F1} and Q^2_{F2}, though this is less evident for CCC.

Looking at SI4b and SI4d, it is interesting to note that the highest discrepancy in the smallest data sets (24 and 48 elements) is found for r^2_m-ExPy with the 1/2 prediction set proportion and Q^2_{F2} with the 1/8 prediction set proportion,

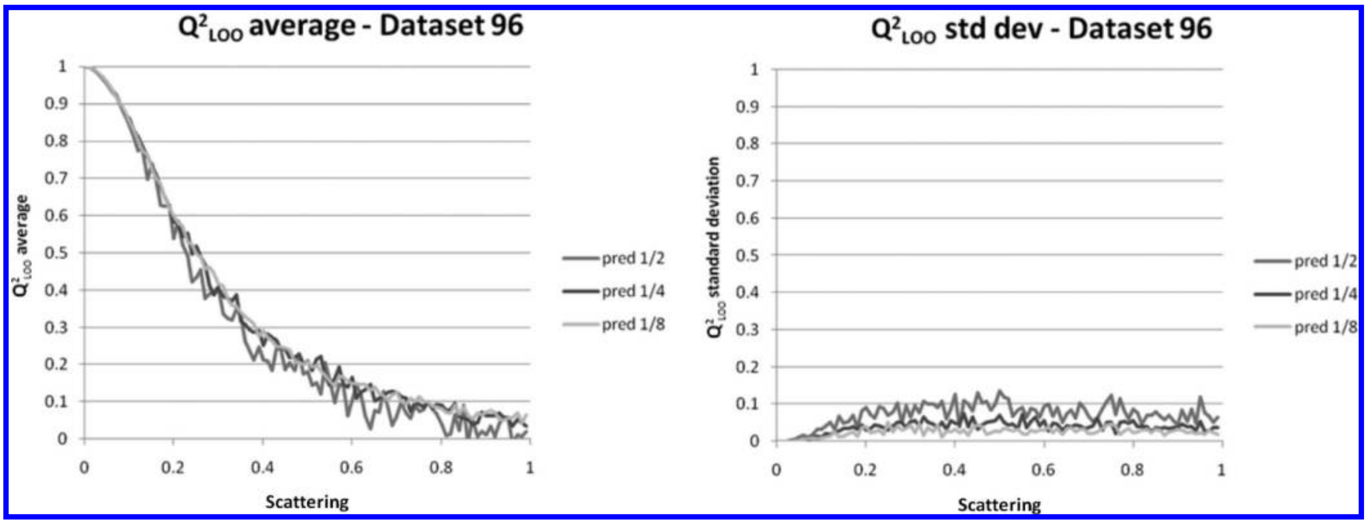


Figure 7. Example of the regularity of the Q^2_{LOO} values in relation to different prediction set proportions.

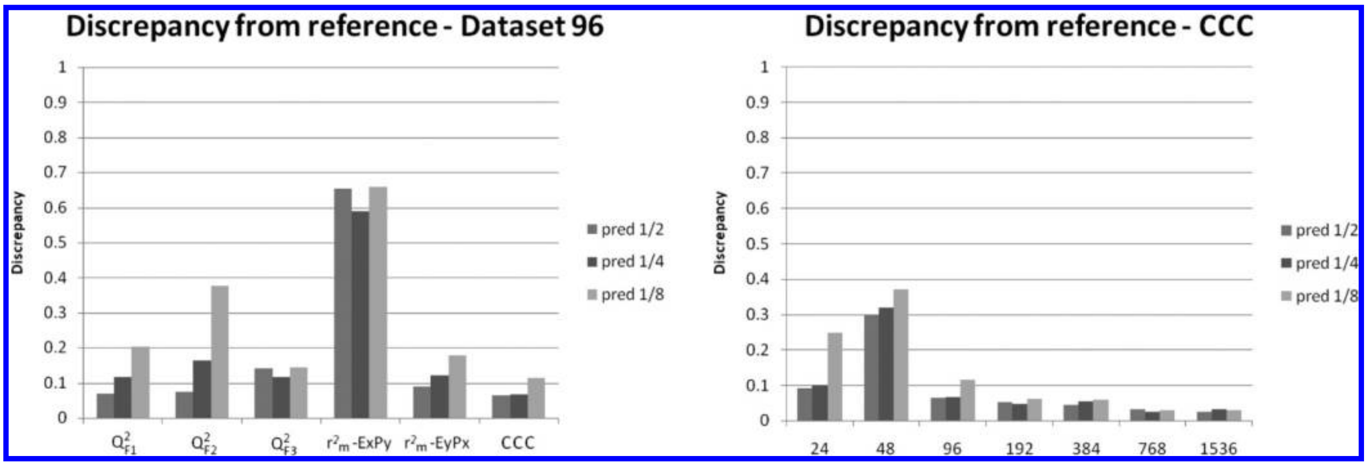


Figure 8. Example of the external validation criteria discrepancies from the corresponding reference values (reported in Figure SI4a) at different prediction set proportion. Left graph: comparison among the various criteria. Right graph: trend of CCC in relation to the data set size.

while the performances relative to the other data set sizes are broadly the same for all the prediction set proportions. Overall, excluding r^2_m-ExPy, the other graphs are qualitatively comparable, but a closer look reveals differences (details in Figure SI4c and SI4e). These figures reveal that the smaller the data set size the better the performances of r^2_m-EyPx and CCC, compared to the other external validation measures. It is important to note that the performance of each criterion is analyzed separately (see Figure SI4d/e), as some of their analytical forms are very different and, as a consequence, it is not possible to make a quantitative comparison of them. Thus, the less the variation in each criterion performance in relation to the data set size (more regular trend), the better the criterion.

Detailed Study of Some Models Extracted from the Simulated Data Sets. To study in greater detail the behavior of all the external validation criteria, in relation to data point scattering and distribution in the graph of experimental versus predicted values, some data sets were selected from the simulated ones (Supporting Information Figures SI5 and SI6a-e).

A peculiar example of the very rare situations in case 1 (see Figure 3), in which CCC accepts models rejected by one or more of the other criteria, is shown in Supporting Information Figure SI5: Q^2_{F1} and Q^2_{F2} reject the model (even if with values very near to the threshold, respectively 0.59 and 0.58), while CCC accepts it (0.87). As it can be noted, the points are almost perfectly aligned on the 45° line and the R^2_{ext} value is 0.998, suggesting that CCC well reflects the data point distribution over the graph.

In the opposite scenario (case 2, Figure 4), where CCC rejects a model while one or more of the other validation measures accept it, many examples are available, thus several data sets of the same size and prediction set proportion can be found. In this case we selected data sets of 192 elements with a prediction set proportion of 1/4 as a compromise between the data number and the validation measures stability.

The first case (Supporting Information Figure SI6a) concerns data points, somewhat dispersed, that lie around the slope 1 line passing through the origin. This model, accepted by all the other external validation measures (both axes dispositions of r^2_m are considered), is rejected by CCC (0.78), as it requires that the data points are nearer the diagonal (i.e., the experimental vs the calculated values should be more similar). Looking at Supporting Information Figure SI6b, r^2_m-ExPy accepts this model, but CCC strongly rejects it, as well as the other criteria.

Looking at Supporting Information Figure SI6c, all data points are almost aligned and are systematically shifted downward: this model is rejected by CCC, while two other external validation measures accept it ($Q^2_{F1} = 0.67$ and both r^2_m, 0.79−0.83). It can be concluded that CCC usually rejects models derived from data too downward or upward biased, even if almost aligned. A similar scenario, but just with more scattered data (see Supporting Information Figure SI6d where R^2 value is 0.92 instead of 0.99), shows more discordant conclusions among the validation criteria: CCC (0.80) rejects the model, while Q^2_{F1} (0.65), Q^2_{F3} (0.60), both r^2_m (0.81 and 0.71), and the Golbraikh and Tropsha method accept it. Due to such dissimilarities among the validation measures, this scenario is further studied on various examples with similar plottings (Supporting Information Figure SI6e) to verify whether the validation criteria that accept this kind of models are always the same. As can be verified, the external validation criteria are not the same in accepting or rejecting models in all similar situations. From Figure SI6b-e, it seems that r^2_m, that accept all the models, does not take into account the slope of the regression lines (both passing, or not, through the origin), so the data can be in a good linear relationship, but the predicted data are not similar to the experimental values. On the contrary, in such a scenario, CCC is always able to better recognize such crucial situations.

Study of Some Examples Taken from Real Data Sets. Some models of mutagenicity of nitro-PAH, developed by our group,[37] were presented[6] as examples of models internally robust (high Q^2_{LOO} values) but externally not predictive (verified by Q^2_{F1}), and these are further explored, in the present work, in order to compare all the external validation criteria presented: the results are shown in Table 1, the corresponding graphs of experimental vs predicted data in Figure 9 and Supporting Information Figure SI-7a to SI-7e.

As can be noted in Table 1, the majority of results are discordant regarding model predictivity, e.g. r^2_m-ExPy and Q^2_{F3} accept the model ID 7 (SIC2 BELv5), while the other criteria reject it. From the predictions in the data point graph of Figure 9 (left graph above) it is evident that rejection would be the best choice, a decision supported by CCC. A similar case is the ID-1 model (PW2 IC1), where the external validation data plotting is very similar (Figure 9, right graph above): it is interesting to note that CCC changed very little (0.79 instead of 0.78) as Q^2_{F1-F3}, while r^2_m and G&T change to a slightly higher extent. In particular, the Golbraikh and Tropsha method accepts it, but just because the slope in the first case was 0.84, just a little below the acceptance threshold, while in the second case it was 0.89, just above the threshold. All the other criteria taken into account in the G&T method were within the acceptance values.

On the other hand, there are examples of coherence among all the external validation measures in rejecting the models ID 3 (VED2 R6u+) (Supporting Information Figure SI-7b) and ID 4 (HATS3u R3v) (Supporting Information Figure SI-7c) (probably because of the strong outlier that can be noted on the abscissa axis).

Using another set of real data (modeling of boiling points of perfluorinated compounds[33]) and looking for problematic models, there are new examples of models to be commented on (Figure 9; graphs down): CCC (the same value for both models) rejects both the models, while the other validation criteria show discordance in the conclusions for predictivity and all accept only the model of the left graph. Note that Q^2_{F1}, Q^2_{F2}, and r^2_m accept both models, while Q^2_{F3} and the G&T method are concordant in accepting the model on the left and rejecting the model on the right. A more cautionary approach should be taken, thus the best choice, which should also be derived from the graph view, would be to reject both models, as suggested by CCC.

Concerning the Use of RMSEP. As previously reported, both relative and absolute measures of errors are needed for a correct assessment of model predictivity. An example is a log P model with a high RMSEP value (e.g., 1) and a high value of the relative measure (e.g., $Q^2_{F1} = 0.8$). The relative measure alone can be misleading, and, in fact, the model must be rejected as not well predictive due to the high absolute error measure, even though the relative measure would accept it. On the other hand, there are cases where the RMSEP (the absolute error measure) alone does not give a correct picture, if the data ranges are not taken into account, as reported in Example SI1.

The relative error measure can be calculated easily by an automatic procedure in the first assessment step of model predictivity, but this is not sufficient and the second step of calculating the absolute error measure (thus also taking into account the scale of measures) is always necessary.

We studied the respective trends of the CCC values, in relation to the RMSEP values, for all the simulated data sets. As expected, the bigger the RMSEP values the smaller the corresponding CCC, as reported in Figure SI8. The use of both CCC and RMSEP values can be considered as a minimal, but sufficient, set

Table 1. Models (Nitro-PAH Mutagenicity[6,37]) with Discordant External Validation Criteria Values[a]

ID	variables	R^2	Q^2_{LOO}	Q^2_{F1}	Q^2_{F2}	Q^2_{F3}	CCC	r^2_m-ExPy	r^2_m-EyPx	G&T
1	PW2 IC1	0.81	0.77	0.60	0.53	0.73	0.79	0.63	0.53	accepted
2	BELe8 HATS4u	0.81	0.76	0.48	0.38	0.65	0.73	0.52	0.46	accepted
3	VED2 R6u+	0.79	0.76	0.27	0.14	0.51	0.58	0.26	0.28	rejected
4	HATS3u R3v	0.80	0.76	0.00	0.00	0.00	0.39	0.17	0.11	rejected
5	BELe8 R4u+	0.80	0.75	0.50	0.42	0.67	0.74	0.55	0.46	accepted
6	SIC2 BEHm8	0.79	0.75	0.61	0.55	0.74	0.8	0.6	0.51	accepted
7	SIC2 BELv5	0.79	0.75	0.58	0.51	0.72	0.78	0.58	0.48	rejected

[a] CCC = concordance correlation coefficient, r^2_m-ExPy = experimental values on the abscissa axis, r^2_m-EyPx = experimental values on the ordinate axis, G&T = Golbraikh and Tropsha method.

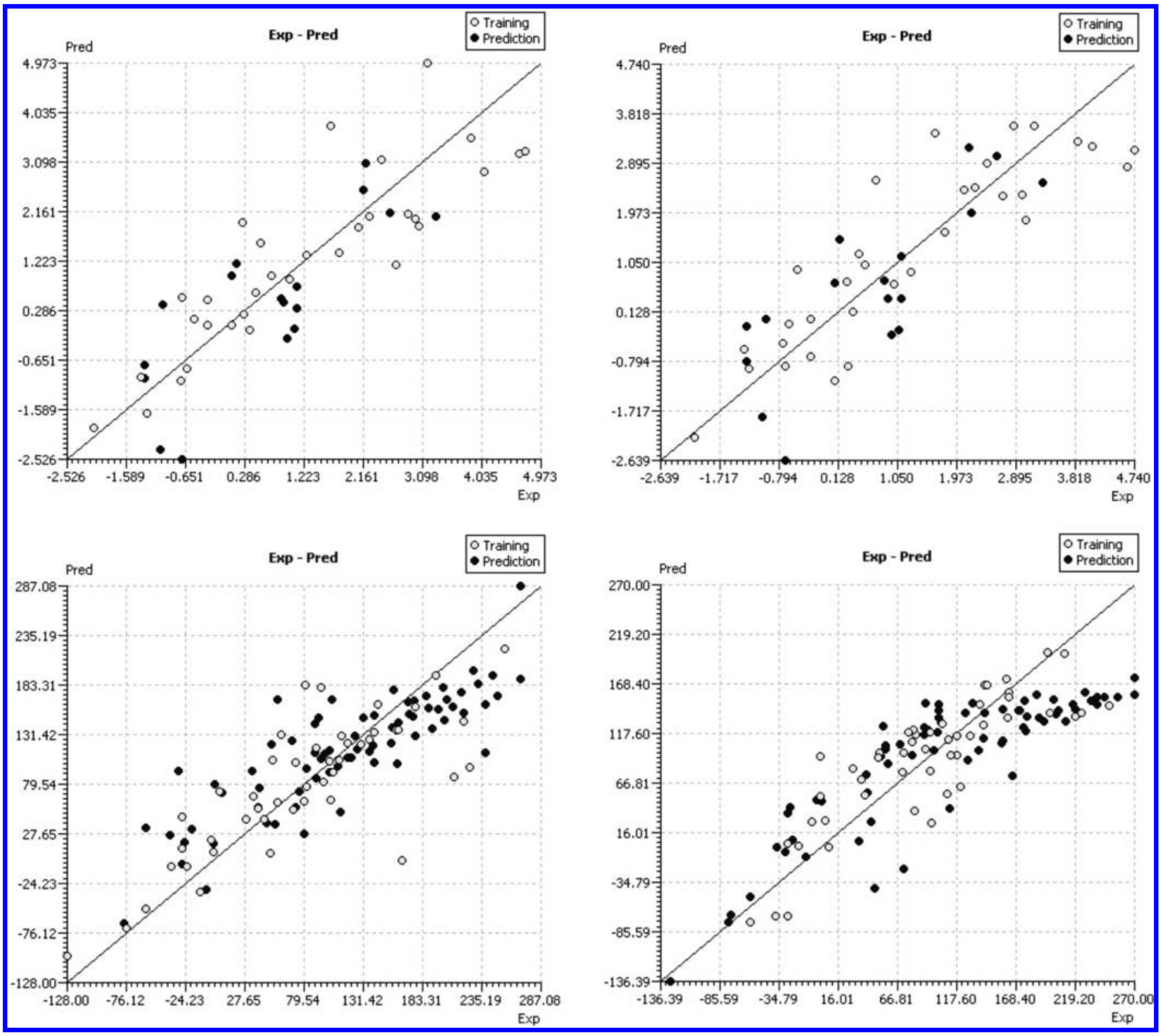


Figure 9. Problematic models from a real data set (Nitro-PAH mutagenicity[36,37]): model ID 7 (SIC2 BELv5, left graph above) and ID-1 (PW2 IC1, right graph up); see Table 1 for external validation criteria values. Other examples from boiling points of perfluorinated compounds[33] (left and right graphs down): left graph external validation criteria values: Q^2_{F1} = 0.74, Q^2_{F2} = 0.69, Q^2_{F3} = 0.67, G&T = accept, r^2_m-ExPy = 0.40, r^2_m-EyPx = 0.63, CCC = 0.79. Right graph values: Q^2_{F1} = 0.69, Q^2_{F2} = 0.68, Q^2_{F3} = 0.47, G&T = reject, r^2_m-ExPy = 0.53, r^2_m-EyPx = 0.71, CCC = 0.79.

of criteria to evaluate a model as externally predictive, in a relative and absolute sense.

CONCLUSIONS

A conceptually simple statistical parameter, the concordance correlation coefficient (CCC), is proposed as a criterion for the external validation of a QSAR model for predictivity on new chemicals. Its performance has been compared to other external validation measures proposed in the literature, using threshold values defined by the proponent, where available, or those normally applied in current practice by QSAR modelers. We verified the criteria performance by simulating a large number of big data sets, as it could be misleading, and perhaps even hazardous, to propose general guidelines to check QSAR model predictive quality based on the results of only a few examples. In this simulation CCC was found to be the most restrictive criterion in accepting a model as externally predictive.

Moreover, after analyzing 210,000 data sets of realistic size, it was verified that the CCC performance is in a good agreement with the other validation measures (ca. 96%) in realistic situations, i.e. when no shifts in the responses are added. In the remaining situations where the validation measures are discordant, it was also demonstrated that in almost all cases it is the most restrictive. In rare cases, related only to the smallest data sets, it is less restrictive. The least restrictive, and thus the most optimistic validation criterion is always r^2_m, followed by Q^2_{F1}, Q^2_{F2}, Q^2_{F3}, and the Golbraikh and Tropsha method.

In addition, we compared the CCC performance with the other validation criteria, in terms of discrepancy from a reference and trend in relation to the data set size: it proved to have a good stability. Additionally, we demonstrated that, as expected, as data scattering increases the CCC trend decreases, while RMSEP increases.

Using real data sets, there are examples where, in the graphs plotting the predicted vs the experimental values, the points distributions are very similar, and, accordingly, CCC almost gives the same values, while the other external validation criteria change to a greater extent. In such situations, our proposed criterion proved to be more stable. In cases where the other external validation measures conflict, a reliably predictive model could be the one that is accepted by all the calculated criteria values (our in-home software[42] always checks this point). Indeed, CCC can help to make the final decision of whether a model should be accepted as predictive or not.

Thus we propose the concordance correlation coefficient as the single validation measure (also because it is based on a simple concept) to evaluate the real predictivity of QSAR models, in accordance with a precautionary approach. In fact, CCC is the most restrictive in accepting models for their external prediction ability, and the separate use of some of the other criteria, with the commonly applied thresholds, could be less precautionary, resulting in the acceptance of not predictive models that are optimistically defined as externally predictive, this is particularly the case of r^2_{m}.

ASSOCIATED CONTENT

Supporting Information. Formula SI1: concordance correlation coefficient formula rearrangement. Figure SI1a: validation criteria performances adding a systematic shift of 0.1. Figure SI1b: detail on Q^2_{F2} and Q^2_{F3} of Figure SI1a. Figure SI1c: validation criteria performances adding a systematic shift of 0.2. Figure SI1d: validation criteria performances adding a systematic shift of 0.6. Figure SI1e: detail on r^2_{m} values of Figure SI1d. Figure SI2a to SI2g: averages and standard deviations of the external validation criteria calculated over ten rounds of simulated data sets using the entire scattering range with no systematic shift added to the responses. Figure SI3: comparative graphs of Q^2_{LOO} values (average over 10 rounds) for every simulated data set size. Figure SI4a: external validation criteria reference calculated averaging 100 simulated data sets of 2000 elements. Figure SI4b: external validation criteria discrepancies calculated with respect to the reference. Figure SI4c: details of Figure SI4b scaled over of the smallest discrepancy values. Figure SI4d: external validation criteria discrepancies, organized by data set size, calculated with respect to the reference. Figure SI4e: details of Figure SI4d scaled over the smallest discrepancy values. Figure SI5: example of studied case 1. Figure SI6a-e: details on some examples of studied case 2. Figure SI7a-e: plotting of the predictions of the nitro-PAH mutagenicity[6,37] real data set. Figure SI8: RMSE and CCC comparison. Example SI1: RMSEP dependence on scale measures. Table SI1: total number of validation criteria that accept the model vs acceptance or rejection by the concordance correlation coefficient. Table SI2: detailed view of Table SI1. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

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

ACKNOWLEDGMENT

The authors thank Dr. Ester Papa, Ph.D. and Prof. Giorgio Binelli for helpful discussions. The anonymous reviewers are thanked for helpful suggestions to improve the manuscript. 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) Kubinyi, H. From narcosis to Hyperspace: The History of QSAR. *Quant. Struct-Act. Relat.* **2002**, *21*, 348–356.

(3) Baumann, K Cross-validation as the objective function for variable-selection techniques. *Trends Anal. Chem.* **2003**, *22*, 395–406.

(4) 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.

(5) Baumann, K.; Stiefl, N. Validation tools for variable subset regression. *J. Comput.-Aided Mol. Des.* **2004**, *18*, 549–562.

(6) Gramatica, P. Principles of QSAR models validation: internal and external. *QSAR Comb. Sci.* **2007**, *5*, 694–701.

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

(8) Kiralj, R.; Ferreira, M. M. C. Basic Validation Procedures for Regression Models in QSAR and QSPR Studies: Theory and Application. *J. Braz. Chem. Soc.* **2009**, *20*, 770–787.

(9) Tropsha, A. Best Practices for QSAR Model Development, Validation, and Exploitation. *Mol. Inf.* **2010**, *29*, 476–488.

(10) Hawkins, D. M. The problem of overfitting. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1–12.

(11) Hawkins, D. M.; Kraker, J. J.; Basak, S. C.; Mills, D. QSPR checking and validation: a case study with hydroxyl radical reaction rate constant. *SAR QSAR Environ. Res.* **2008**, *16*, 525–539.

(12) Helma, C. Data mining and knowledge discovery in predictive toxicology. *SAR QSAR Environ. Res.* **2004**, *15*, 367–383.

(13) Jensen, G. E.; Niemelä, J.R.; Wedebye, E. B.; Nikolov, N. G. QSAR models for reproductive toxicity and endocrine disruption in regulatory use — a preliminary investigation. *SAR QSAR Environ. Res.* **2008**, *19*, 631–641.

(14) Golbraikh, A.; Shen, M.; Xiao, Z. Y.; Xiao, Y. D.; Lee, K. H.; Tropsha, A. Rational selection of training and test sets for the development of validated QSAR models. *J. Comput.-Aided Mol. Des.* **2003**, *17*, 241–253.

(15) Gramatica, P.; Pilutti, P.; Papa, E. Validated QSAR Prediction of OH Tropospheric degradability: splitting into training-test set and consensus modeling. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1794–1802.

(16) Leonard, J. T.; Roy, K. On Selection of Training and Test Sets for the Development of Predictive QSAR models. *QSAR Comb. Sci.* **2006**, *3*, 235–251.

(17) Bhhatarai, B.; Gramatica, P. Per- and Poly-fluoro Toxicity (LC50 inhalation) Study in Rat and Mouse using QSAR Modeling. *Chem. Res. Toxicol.* **2010**, *23*, 528–539.

(18) 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.

(19) 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.

(20) 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.

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

(22) Roy, P. P.; Roy, K. On Some Aspects of Variable Selection for Partial Least Squares Regression Models. *QSAR Comb. Sci.* **2008**, *27*, 302–313.

(23) Roy, P. P.; Somnath, P.; Mitra, I.; Roy, K. On two novel parameters for validation of predictive QSAR models. *Molecules* **2009**, *14*, 1660–1701.

(24) Mitra, I.; Roy, P. P.; Kar, S.; Ojha, P. K.; Roy, K. On further application of r^2_m as a metric for validation of QSAR models. *J. Chemom.* **2010**, *24*, 22–33.

(25) 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.

(26) Todeschini, R.; Consonni, V.; Pavan, M. *MOBY DIGS*, version 1.2; Software for multilinear regression analysis and Variable subset selection by genetic algorithm; Talete srl: Milan, Italy, 2002.

(27) Organization for Economic Co-operation and Development (OECD). Guidance document on the validation of (quantitative) structure-activity relationship [(Q)SAR] models, 2007. OECD Web Site. http://www.oecd.org/officialdocuments/displaydocumentpdf/?cote=env/jm/mono%282007%292&doclanguage=en (accessed July 18, 2011).

(28) 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.

(29) Bhhatarai, B.; Garg, R.; Gramatica, P. Are mechanistic and statistical QSAR approaches really different? MLR studies on 158 cycloalkyl-pyranones. *Mol. Inf.* **2010**, *29*, 511–522.

(30) Li, J.; Gramatica, P. The importance of molecular structures, endpoints' values and predictivity parameters in QSAR research - QSAR analysis of a series of estrogen receptor binders. *Mol. Diversity* **2010**, *14*, 687–696.

(31) Bhhatarai, B.; Gramatica, P. Oral LD_{50} Toxicity Modeling and Prediction of Per- and Polyfluorinated Chemicals on Rat and Mouse. *Mol. Diversity* **2011**, *15*, 467–476.

(32) Bhhatarai, B.; Gramatica, P. Predicting physico-chemical properties of emerging pollutants: QSPR modeling of Benzo(triazoles). *Water Res.* **2011**, *45*, 1463–1471.

(33) Bhhatarai, B.; Teetz, W.; Liu, T.; Oberg, T; Jeliazkova, N.; Kochev, N.; Pukalov, O.; Tetko, I.; Kovarich, S.; Papa, E.; Gramatica, P. CADASTER QSPR Models for Predictions of Melting and Boiling Points of Perfluorinated Chemicals. *Mol. Inf.* **2011**, *30*, 189–204.

(34) Lin, L. I. A Concordance Correlation Coefficient to Evaluate Reproducibility. *Biometrics* **1989**, *45*, 255–268.

(35) Lin, L. I. Assay Validation Using the Concordance Correlation Coefficient. *Biometrics* **1992**, *48*, 599–604.

(36) Benigni, R. QSARs for Mutagenicity and Carcinogenicity. In The report from the expert group on (quantitative) structure-activity relationships [(Q)SARs] on the principles for the validation of (Q)SARs, 2004, pp 84−100. OECD Web Site. http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=ENV/JM/MONO-(2004)24&docLanguage=En (accessed July 18, 2011).

(37) Gramatica, P.; Pilutti, P.; Papa, E. Approaches for externally validated QSAR modeling of Nitrated Polycyclic Aromatic Hydrocarbon mutagenicity. *SAR QSAR Environ. Res.* **2007**, *18*, 169–178.

(38) ChemIDplus Advanced. http://chem.sis.nlm.nih.gov/chemidplus (accessed July 18, 2011).

(39) Hendricks, J. O. Industrial fluoro-chemicals. *Ind. Eng. Chem.* **1953**, *45*, 99–105.

(40) Roy, P. P.; Leonard, T. J.; Roy, K. Exploring the impact of size of training sets for the development of predictive QSAR models. *Chemom. Intell. Lab. Syst.* **2008**, *90*, 31–42.

(41) Faber, N. M. Estimating the uncertainty in estimates of root mean square error of prediction: application to determining the size of an adequate test set in multivariate calibration. *Chemom. Intell. Lab. Syst.* **1999**, *49*, 79–89.

(42) Chirico, N.; Papa, E.; Kovarich, S.; Cassani, S.; Gramatica, P. *QSARINS, software for QSAR model calculation and validation, under development*; University of Insubria, Varese, Italy, 2011.