Journal of
CHEMICAL INFORMATION
AND MODELING

Article

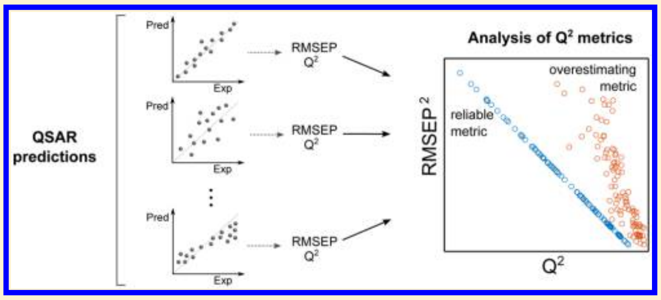
pubs.acs.org/jcim

Beware of Unreliable Q^2! A Comparative Study of Regression Metrics for Predictivity Assessment of QSAR Models

Roberto Todeschini,* Davide Ballabio, and Francesca Grisoni

Milano Chemometrics and QSAR Research Group, Department of Earth and Environmental Sciences, University of Milano-Bicocca, P.za della Scienza, 1, 20126 Milan, Italy

ABSTRACT: Validation is an essential step of QSAR modeling, and it can be performed by both internal validation techniques (e.g., cross-validation, bootstrap) or by an external set of test objects, that is, objects not used for model development and/or optimization. The evaluation of model predictive ability is then completed by comparing experimental and predicted values of test molecules. When dealing with quantitative QSAR models, validation results are generally expressed in terms of Q^2 metrics. In this work, four fundamental mathematical principles, which should be respected by any Q^2 metric, are introduced. Then, the behavior of five different metrics (Q^2_{F1}, Q^2_{F2}, Q^2_{F3}, Q^2_{CCC}, and Q^2_{Rm}) is compared and critically discussed. The conclusions highlight that only the Q^2_{F3} metric satisfies all the stated conditions, while the remaining metrics show different theoretical flaws.



1. INTRODUCTION

An essential step of chemical modeling is to evaluate the model predictive ability in an objective manner. This is generally performed by evaluating the predictions on molecules that were not used to build/calibrate the model, i.e. by using (1) some training set molecules, through internal validation techniques (e.g., cross-validation, bootstrapping[1]), or (2) an external set of molecules not used for model calibration. In both the cases, this is performed by comparing the predictions on test data with the actual values of the property to model. Thus, a reliable estimate of the model predictive ability toward new data is necessary to test for the presence of overfitting, model instability, or other pathologies, and to ensure the optimal model applicability.[1−3]

The validation plays a central role in many chemoinformatic and modeling applications,[4,5] and this is particularly true for quantitative structure−activity relationship (QSAR) modeling,[6−8] which aims to define a mathematical relationship between molecular properties (encoded within the so-called molecular descriptors[9]) and a biological property of interest. In the field of QSAR, the external validation is regarded as a necessary step for the objective model evaluation, in particular for regulatory applications, as stated by the Organization for Economic Cooperation and Development.[10]

Assessment of model predictivity is usually expressed through the root mean squared error in prediction (*RMSEP*), also known as the standard deviation error in prediction (*SDEP*). The *RMSEP* is a measure of the mean model error on new data (the higher the *RMSEP*, the higher the error) and is expressed in the same measuring unit of the modeled biological property. This aspect hampers the possibility to directly compare: (a) models calibrated on responses with different measuring units, and (b) models calibrated on different properties. To this end, throughout the years, several metrics (Q^2) have been proposed to quantify the predictive ability of models, independently from the measuring units of the response.[11−17] The purpose of any Q^2 metric is to transform the information encoded by the *RMSEP* into an upper bounded index, ranging from −∞ to 1: the higher the Q^2 metric, the higher the predictivity (and the smaller the model error). Therefore, a proper Q^2 metric should be invariant to the response scale and closely (and inversely) related to *RMSEP*. Obviously, the optimal metric is not the one providing the highest/lowest values, but the one that better quantifies the true model predictivity.

Stemming from these considerations, this work extends two of our previous studies investigating three Q^2 metrics[11,12] and it: (1) includes two more recent parameters,[13−16] and (2) broadens the set of evaluation criteria. In particular, in this work, four rigorous mathematical requirements were formalized for evaluating the Q^2 metrics, namely: (1) invariance to the response scale, (2) invariance for a fixed *RMSEP* value, (3) correlation with *RMSEP*, and (4) invariance to the splitting of the external set in subsets (ergodic principle). These principles are the basic requirements for any metric to be representative of the true predictivity of a model toward external test molecules and were the starting point of the analysis.

In this study, after introducing the rationale of the mathematical principles and the formulations of the Q^2 metrics, the traits of the analyzed metrics are highlighted and discussed, using *ad hoc* simulated data sets. The study identified only one optimal metric out of five, and underscored that, in most of the cases, the other metrics are suboptimal and affected by several types of flaws.

Received: May 16, 2016
Published: September 15, 2016

DOI: 10.1021/acs.jcim.6b00277
J. Chem. Inf. Model. 2016, 56, 1905−1913

2. THEORY

2.1. Benchmark Measures. In regression analysis, the traditional quantity to estimate the model quality is the root mean squared error in calculation (*RMSEC*), calculated as the square root of the average of the residual sum of squares (*RSS*):

$$RMSEC = \sqrt{\frac{\sum_{i=1}^{n_{TR}} (y_i - \hat{y}_i)^2}{n_{TR}}} = \sqrt{\frac{RSS}{n_{TR}}} \tag{1}$$

where n_{TR} is the number of objects included in the training set (i.e., used to calibrate the regression model), while y_i and $\hat{y}_i$ are the experimental and the calculated response of the *i*-th object, respectively. Since the *RMSEC* depends on the scale of the response measure, the classical measure of the model quality, independent from the response scale, is the coefficient of determination (R^2), representing the variance explained by the model, as follows:

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{i=1}^{n_{TR}} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n_{TR}} (y_i - \bar{y}_{TR})^2} \tag{2}$$

where $\bar{y}_{TR}$ is the average response of the training objects, while y_i and $\hat{y}_i$ are the experimental and the calculated response of the *i*-th object, respectively.

Once a model is calculated using the training set objects, its prediction ability is estimated through the root mean squared error in prediction (*RMSEP*), defined as

$$RMSEP = \sqrt{\frac{\sum_{i=1}^{n_{OUT}} (y_i - \hat{y}_{i/i})^2}{n_{OUT}}} = \sqrt{\frac{PRESS}{n_{OUT}}} \tag{3}$$

where $\hat{y}_{i/i}$ is the predicted response for the *i*-th object when it is not present in the training set and n_{OUT} is the number of objects that are not considered in model building; either they are left out by a validation procedure or left out because they belong to an external set. The *RMSEC* is minimized during the least-squares procedure, while the *RMSEP* is accepted by all the researchers as the optimal parameter for evaluating the predictive ability of a model. However, as for *RMSEC*, the *RMSEP* measure also depends on the scale of the response, thus having different scales for different response scales.

As the *RMSEP* is the natural reference measure for evaluating the prediction ability of a regression model,[7] it was the starting point to define some basic principles that should be satisfied by any Q^2 metric. In particular, four principles were formalized and can be found in the following paragraph.

2.2. Mathematical Principles. Hereafter, the following notation will be used: Q^2_M denotes a set of *M* values of a Q^2 metric and $RMSEP_M$ the set of the corresponding root mean square errors.

This paragraph will introduce four basic principles, which are founded on the fact that the *RMSEP* is the reference measure of the model predictivity and that any Q^2 metric should be related to *RMSEP* independently of the measuring unit of the response. The proposed principles are regarded as the most rational mathematical properties of any Q^2 metric, also in relationship with the corresponding *RMSEP*. These properties have to be fulfilled in order to avoid biased or doubtful evaluations of the model predictive power.

2.2.1. Principle 0: Invariance to Response Scaling. As the basic aim of any Q^2 metric is to give a quality measure independent from the response scale, the essential principle (*P0*) is the metric invariance to any linear transformation of the response, which can be formulated as

$$Q^2(Y') = Q^2(Y) \quad \text{where } Y' = a\cdot Y + b \tag{4}$$

2.2.2. Principle 1: Invariance to RMSEP. Assuming *RMSEP* as the reference measure of regression predictive ability, if the $RMSEP_M$ set comprises *M* constant values, then the corresponding Q^2_M set must comprise *M* constant values, that is

$$Q_m^2 = Q_{m'}^2 \quad \text{if } SDEP_m = SDEP_{m'} \quad m, m' = 1, 2, ..., M \tag{5}$$

In other words, Principle 1 implies a null variance of the Q^2 values (i.e., $V(Q^2_M) = 0$) for a set of constant *RMSEP* values.

2.2.3. Principle 2: Correlation to RMSEP. Assuming *RMSEP* as the reference measure of regression predictive ability, any set of *M* values of a regression metric Q^2 must satisfy the following relationship with the corresponding set $RMSEP_M$:

$$\rho(Q_M^2, RMSEP_M^2) = -1 \quad \text{if } RMSEP_m \neq RMSEP_{m'}$$
$$m, m' = 1, 2, ..., M \tag{6}$$

where ρ is the Pearson correlation coefficient.

Principle 2 implies that the *RMSEP* values and the corresponding Q^2 values should be closely and inversely correlated.

It can be noted that Principle 2 implies a correlation equal to −1 between Q^2 and squared *RMSEP* values; otherwise, the relationship between *RMSEP* and Q^2 leads to a Pearson linear correlation close, but not equal, to −1.

2.2.4. Principle 3: Ergodic Principle. In general, a system *S* fulfils the ergodic principle if a given property calculated on the whole system coincides with the average value of the same property calculated on each member S_g of a partition of the system into *G* parts in such a way that

$$S = \cup_g S_g, \quad \bar{P} = \frac{\sum_{g=1}^{G} P_g}{G} \text{ and } P_0 = \bar{P} \tag{7}$$

where P_0 is the property calculated on the whole system, *G* is the number of the system partitions, S_g is the *g*-th partition of the system, and P_g is the property value of the *g*-th partition.

As the prediction of each external object should be independent of all other external objects or of how they are partitioned into sets, the ergodic property should also hold for any Q^2 metric.[11] In other words, any metric calculated over all the test objects (Q^2_0) must coincide with the average $\bar{Q}^2$ metric estimated over the Q^2_g values obtained by any partition of the test objects into *G* groups, as follows:

$$Q_0^2 = \bar{Q}^2 = \frac{\sum_{g=1}^{G} Q_g^2}{G} \tag{8}$$

2.3. Definitions of Q^2 Metrics. In this paragraph, the analyzed Q^2 metrics are briefly introduced. The first three metrics (Q^2_{F1}, Q^2_{F2}, and Q^2_{F3}) were studied and discussed in two of our previous papers;[11,12] they are defined as the following:

$$Q_{F1}^2 = 1 - \frac{\sum_{i=1}^{n_{OUT}} (y_i - \hat{y}_{i/i})^2}{\sum_{i=1}^{n_{OUT}} (y_i - \bar{y}_{TR})^2} \tag{9.1}$$

$$Q_{F2}^2 = 1 - \frac{\sum_{i=1}^{n_{OUT}} (y_i - \hat{y}_{i/i})^2}{\sum_{i=1}^{n_{OUT}} (y_i - \bar{y}_{OUT})^2} \tag{9.2}$$

$$Q_{F3}^2 = 1 - \frac{\sum_{i=1}^{n_{OUT}} (y_i - \hat{y}_{i/i})^2 / n_{OUT}}{\sum_{i=1}^{n_{TR}} (y_i - \bar{y}_{TR})^2 / n_{TR}} \tag{9.3}$$

where y_i is the experimental response of the i-th object, $\hat{y}_{i/i}$ is the predicted response when the i-th object is not in the training set, n_{TR} and n_{OUT} are the number of training and test objects, respectively, $\bar{y}_{TR}$ is the average value of the training set experimental responses, and $\bar{y}_{OUT}$ is the average value of experimental responses of external test objects. It can also be noted that, for classical cross-validation techniques (e.g., leave-one out), where each training object is used as test object only once, $Q_{F1}^2 = Q_{F2}^2 = Q_{F3}^2$, with $n_{OUT} = n_{TR}$ and $\bar{y}_{OUT} = \bar{y}_{TR}$. Note that the metric Q_{F1}^2 is the most used in QSAR modeling.

In the last few years, other metrics for evaluating the prediction ability of external sets were proposed.

The concordance correlation coefficient (*CCC*) was proposed to assess the agreement of two trials of an assay or instrument in terms of method validity and repeatability.[18,19] Later, the concordance correlation coefficient *CCC* was introduced to evaluate the prediction ability of regression models.[13] In particular, the concordance correlation coefficient, here referred to as Q_{CCC}^2, is defined as

$$Q_{CCC}^2 = \left[2\cdot\sum_{i=1}^{n_{OUT}} (y_i - \bar{y}_{EXP})\cdot(\hat{y}_{i/i} - \bar{y}_{PRED})\right] \Big/ \left[\sum_{i=1}^{n_{OUT}} (y_i - \bar{y}_{EXP})^2 + \sum_{i=1}^{n_{OUT}} (\hat{y}_{i/i} - \bar{y}_{PRED})^2 + n_{OUT}\cdot(\bar{y}_{EXP} - \bar{y}_{PRED})^2\right] \tag{10}$$

where $\bar{y}_{EXP}$ and $\bar{y}_{PRED}$ are the average of the experimental and predicted responses of the external set, respectively. It can be noted that no information about the training set is used.

Assuming that the experimental responses of the external set are denoted by X and the predicted responses by Y, Q_{CCC}^2 can be expressed as

$$Q_{CCC}^2 = \frac{2\cdot S_{XY}}{S_{XX} + S_{YY} + (\bar{X} - \bar{Y})^2} \tag{11}$$

where S_{XY}, S_{XX}, and S_{YY} are the covariance between X and Y and their variances, respectively.

The R_m metric, here referred to as Q_{Rm}^2, was proposed to evaluate the predictive ability of regression QSAR models.[14−16] Still assuming that the experimental and predicted responses of the external set are respectively denoted by X and Y, Q_{Rm}^2 is defined as

$$\begin{aligned} Q_{Rm}^2 &= \frac{1}{2}\cdot[R^2\cdot(1 - \sqrt{R^2 - R_{0,XY}^2}) + R^2\cdot(1 - \sqrt{R^2 - R_{0,YX}^2})] \\ &= R^2\cdot\left[1 - \frac{\sqrt{R^2 - R_{0,XY}^2} + \sqrt{R^2 - R_{0,YX}^2}}{2}\right] \end{aligned} \tag{12}$$

where R^2 is the coefficient of determination calculated between X and Y, $R_{0,XY}^2$ is the coefficient of determination of predicted values versus experimental values, and $R_{0,YX}^2$ is the coefficient of determination of the experimental values versus predicted values, both calculated for a regression line through the origin.

Finally, the Golbraikh−Tropsha decision rule (*GTR*) was proposed as a multicriteria decision tool.[20] According to *GTR*, in the last improved versions,[7,21] a model is acceptable if it fulfills all the following conditions:

$$\begin{aligned} &1.\ Q^2 > 0.5 \wedge R^2 > 0.6 \\ &2.\ \frac{R^2 - R_{0,XY}^2}{R^2} < 0.1 \wedge 0.90 \le b_{XY} \le 1.1 \\ &3.\ \frac{R^2 - R_{0,YX}^2}{R^2} < 0.1 \wedge 0.90 \le b_{YX} \le 1.1 \\ &4.\ |R_{0,XY}^2 - R_{0,YX}^2| < 0.3 \end{aligned} \tag{13}$$

where R^2 is the correlation coefficient, $R_{0,XY}^2$ and $R_{0,YX}^2$ are the same quantities previously defined; and b_{XY} and b_{YX} are the two slopes of the corresponding regression lines through the origin. Being *GTR* not a continuous metric but a decision tool, it was not taken into account in the comparison analysis.

3. RESULTS AND DISCUSSION

The five Q^2 metrics were evaluated according to each of the basic principles previously introduced. Simulated training and test set responses were generated in order to thoroughly test the compliance of the metrics to each of the principles. Results are discussed on a point-by-point basis, and finally, the usage of thresholds on Q^2 is also addressed.

3.1. Principle 0: Invariance to Response Scaling. The property of invariance to response scaling is the basic property for any Q^2 metric. The invariance to the response scaling is implicit in the mathematical formulations of Q_{F1}^2, Q_{F2}^2, Q_{F3}^2, and Q_{CCC}^2. In fact, the numerators and denominators of the ratios (eq 9) depend in the same way on the response's scale, so that changing the response units has the same effect on both terms and, consequently, leads to the same metric value. Despite the fact that the invariance to response scaling is in this case obvious, Q_{F1}^2, Q_{F2}^2, Q_{F3}^2, and Q_{CCC}^2 were anyhow analyzed for the sake of completeness.

In order to evaluate the invariance of the metrics to linear transformations of the response, the following procedure was used:

(a) randomly generate a set of 100 experimental responses for training objects with a uniform distribution in the range [0,100]; generate the corresponding 100 calculated responses by adding a random error to the experimental training responses; the random error is uniformly distributed in the range $[-E_{TR}, +E_{TR}]$;

(b) randomly generate a set of 25 experimental responses for test objects with a uniform distribution in the range [0, 100]; generate the corresponding 25 predicted responses by adding a random error to the experimental test responses; the random error is uniformly distributed in the range $[-E_{TS}, +E_{TS}]$;

(c) repeat (b) for 100 iterations; in each iteration, the E_{TS} increases so that $E_{TS1} < E_{TS2} < ... < E_{TS100}$;

(d) for each iteration, calculate the corresponding *RMSEP*; it derives that $RMSEP_1 < RMSEP_2 < ... < RMSEP_{100}$;

(e) for each iteration, calculate Q^2 metrics on the 25 test responses; calculate Q^2 metrics on the 25 test responses scaled in the range [0, 1], using the scaling parameters of the training set;

(f) calculate the difference (in percentage) between Q^2 derived from scaled and not scaled responses:

$$diff\,\% = |Q^2(unscaled) - Q^2(scaled)| \times 100$$

(g) among the 100 *diff %* values, take the maximum;

Table 1. Statistics Calculated for Each Fixed *RMSEC* and *RMSEP* Value over 500 Simulated Repetitions

			Q^2_{F1}		Q^2_{F2}		Q^2_{F3}		Q^2_{CCC}		Q^2_{Rm}	
ID	*RMSEC*	*RMSEP*	*mean*	*st.dev*	*mean*	*st.dev*	*mean*	*st.dev*	*mean*	*st.dev*	*mean*	*st.dev*
1	2.688	2.779	0.991	0.002	0.990	0.002	0.992	0.000	0.995	0.001	0.986	0.006
2	5.523	6.901	0.942	0.012	0.938	0.013	0.939	0.000	0.970	0.006	0.934	0.018
3	9.027	8.500	0.910	0.019	0.905	0.022	0.920	0.000	0.955	0.010	0.928	0.021
4	11.940	11.517	0.835	0.033	0.826	0.034	0.845	0.000	0.920	0.016	0.859	0.029
5	13.699	16.561	0.662	0.070	0.637	0.077	0.702	0.000	0.847	0.031	0.785	0.052
6	17.713	18.546	0.575	0.085	0.555	0.094	0.550	0.000	0.817	0.037	0.713	0.064
7	19.580	21.831	0.410	0.117	0.383	0.127	0.477	0.000	0.762	0.049	0.648	0.083
8	24.044	24.388	0.300	0.147	0.222	0.163	0.349	0.000	0.718	0.056	0.594	0.081
9	26.840	28.219	0.025	0.188	−0.031	0.202	0.065	0.000	0.653	0.067	0.528	0.100
10	29.194	29.289	−0.072	0.219	−0.121	0.242	−0.019	0.000	0.642	0.065	0.581	0.099

[a]Mean and standard deviation for each metric are reported.

Table 2. Minimum and Maximum Values of the Five Metrics Calculated for Each Fixed *RMSEC* and *RMSEP* Value (see Table 1) over 500 Simulated Repetitions

	Q^2_{F1}		Q^2_{F2}		Q^2_{F3}		Q^2_{CCC}		Q^2_{Rm}	
ID	*min*	*Max*	*min*	*Max*	*min*	*Max*	*min*	*Max*	*min*	*Max*
1	0.978	0.994	0.978	0.994	0.992	0.992	0.989	0.997	0.959	0.994
2	0.874	0.963	0.865	0.961	0.939	0.939	0.938	0.981	0.872	0.966
3	0.775	0.949	0.765	0.943	0.920	0.920	0.896	0.973	0.841	0.966
4	0.699	0.901	0.683	0.895	0.845	0.845	0.854	0.949	0.730	0.919
5	0.242	0.799	0.240	0.796	0.702	0.702	0.727	0.907	0.604	0.901
6	0.213	0.727	0.057	0.720	0.550	0.550	0.684	0.887	0.484	0.830
7	−0.262	0.630	−0.262	0.610	0.477	0.477	0.552	0.857	0.350	0.792
8	−0.368	0.573	−0.406	0.532	0.349	0.349	0.457	0.828	0.245	0.750
9	−0.886	0.387	−0.995	0.365	0.065	0.065	0.378	0.796	0.189	0.737
10	−1.352	0.343	−1.374	0.339	−0.019	−0.019	0.281	0.787	0.164	0.819

(h) repeat the procedure (a−g) five times;

(i) repeat the procedure (a−h) for 10 iterations; in each iteration, E_{TR} is increased in increments of 5 units from 5 to 50, so that $RMSEC_1 < RMSEC_2 < ... < RMSEC_{10}$.

Four out of five metrics resulted rigorously invariant to the response scaling, with the exception of Q^2_{Rm}. In fact, the maximum differences *diff* % in all the trials for Q^2_{F1}, Q^2_{F2}, Q^2_{F3}, and Q^2_{CCC} were always lower than 10^{-14}, while, on the contrary, Q^2_{Rm}, as reported by its author,[22] is not invariant and showed differences up to 1%.

The origin of this drawback is that the coefficient of determination R^2_0 of a regression line through the origin is not fully invariant to the response scaling. Indeed, for any linear transformation of the experimental response, such as $Y' = aY + b$, the scaling invariance is fulfilled if and only if $b = 0$.

For example, the range scaling *RS* is defined as

$$y_i'(RS) = \frac{y_i - L}{U - L}$$

where U and L are the maximum and minimum values of the experimental responses and the scaling parameters a and b are

$$a = \frac{1}{U - L} \text{ and } b = \frac{-L}{U - L}$$

Thus, the invariance is fulfilled only if the minimum L of the response is equal to zero.

With the data being randomly extracted between [0, 100], the minimum value L is near to zero for all the repetitions; that is, this is the optimal condition to obtain scaling invariance. Despite this favorable condition, Q^2_{Rm} shows some differences: this is a formal drawback for the Q^2_{Rm} metric because it contradicts the aim of a Q^2 metric, i.e. to transform a quantity dependent from the scale of the response (*RMSEP*) into a normalized independent quantity. In order to overcome this drawback, the authors[25] proposed to calculate Q^2_{Rm} after a pretreatment of the response by a range scaling.

For the Gobraikh−Tropsha rule, which also uses the quantity R^2_0, this drawback is not relevant because *GTR* is not defined as a continuous metric but as a decision tool, which is robust to the small differences between scaled and not scaled values of the implied parameters.

3.2. Principle 1: Invariance to RMSEP.

In order to evaluate the metric fulfillment of the first principle, regarding the invariance to the *RMSEP*, the following procedure was applied:

(a) randomly generate a set of 100 experimental responses for training objects with a uniform distribution in the range [0, 100]; generate the corresponding 100 calculated responses by adding a random error to the experimental training responses; the random error is uniformly distributed in the range $[-E_{TR}, +E_{TR}]$;

(b) randomly generate a set of 25 experimental responses for test objects with a uniform distribution in the range [0, 100]; generate the corresponding 25 predicted responses by adding a random error to the experimental test responses; the random error is uniformly distributed in the range $[-E_{TS}, +E_{TS}]$;

(c) repeat (b) for 500 iterations; in each iteration *RMSEP* is maintained constant, so that $RMSEP_1 = RMSEP_2 = ... =$

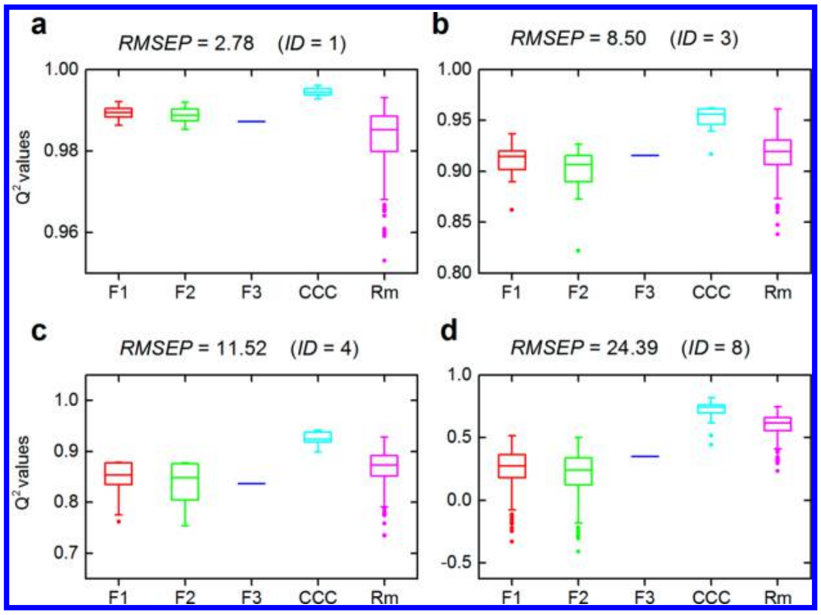


Figure 1. Box-plots of the five Q^2 metrics obtained by 500 repetitions with the same *RMSEP*, for four different *RMSEP* values (a–d).

RMSEP$_{500}$ and calculate the corresponding 500 Q^2 values;

(d) calculate average, minimum, maximum, and the standard deviation over the 500 Q^2 values associated with the same *RMSEP*;

(e) repeat the procedure (a–d) for 10 iterations, with increasing *RMSEC* and *RMSEP*; in each iteration, E_{TR} and E_{TS} increase in increments of 5 units from 5 to 50, so that $RMSEC_1 < RMSEC_2 < ... < RMSEC_{10}$ and $RMSEP_1 < RMSEP_2 < ... < RMSEP_{10}$.

For each trial, the standard deviations and average values calculated over the 500 repetitions are reported (Table 1) along with minimum and maximum values (Table 2). In Figure 1, the distribution of the 500 Q^2 values for each metric is graphically shown for 4 out of the 10 cases.

The rationale of Principle 1 is that a given *RMSEP* value should always reflect in the same value of the corresponding Q^2 metric. Thus, in the case of many models giving the same *RMSEP*, one would expect to have constant values of the associated Q^2 metric. However, this is not the case of Q^2_{F1}, Q^2_{F2}, Q^2_{CCC}, and Q^2_{Rm}, whose standard deviations over the 500 repetitions are always greater than zero and significantly increase with increasing *RMSEP* (Table 1). Only the Q^2_{F3} metric has constant values for constant *RMSEP* values.

Looking at the Q^2 average values, it can be noted that, for high *RMSE* cases (e.g., *IDs* from 7 to 10 in Table 1), Q^2_{CCC} and Q^2_{Rm} are remarkably larger than the other metrics, the former being larger than 0.64 and the latter being larger than 0.52. In other words, both metrics show an awkward tendency to overestimate the predictive ability, in particular for increasing *RMSEP* values, as also highlighted in Figure 1. The overestimation tendency of Q^2_{CCC} was already underscored[23] along with its being very sensitive both to the heterogeneity of objects[24] and to the distribution of test responses around the mean.[25]

The overestimation issue of Q^2_{CCC} and Q^2_{Rm} is related to their being calculated by refitting the experimental versus the predicted responses of the external set, in some way, neglecting the model R^2 on the training set. Correspondingly, the Q^2 values calculated on this "secondary model"[7] are always too optimistic. Moreover, despite some authors stating that neglecting the training set to calculate a Q^2 metric is advantageous,[27] in our opinion, the training set information should always be taken into account for several reasons, such as to be used for benchmarking the mean model error, for evaluating the applicability domain, and for testing the presence of model pathologies.

Moreover, the minimum and maximum values of Q^2_{F1}, Q^2_{F2}, Q^2_{CCC}, and Q^2_{Rm} significantly diverge when *RMSEP* increases (Table 3 and Figure 1). On the contrary, Q^2_{F3} has equal maximum and minimum values, over the 500 repetitions, independently from the *RMSEP* magnitude. For example, in the intermediate *RMSEP* case (*ID* 6, *RMSEP* = 18.546), Q^2_{F1} shows a difference between minimum and maximum greater than 50%, Q^2_{F2} about 66%, Q^2_{CCC} about 20%, and Q^2_{Rm} about 35%; in the largest *RMSEP* case (*ID* 10, *RMSEP* equal to 29.289), Q^2_{CCC} shows a difference greater than 50% and Q^2_{Rm} greater than 65%.

Table 3. Pearson Correlations between Each Metric and the Corresponding *RMSEP*^{2} Values Obtained over 100 Repetitions with Increasing *RMSEC* Values

ID	R^2	*RMSEC*	Q^2_{F1}	Q^2_{F2}	Q^2_{F3}	Q^2_{CCC}	Q^2_{Rm}
1	0.988	3.057	−0.933	−0.934	−1	−0.934	−0.923
2	0.959	5.731	−0.950	−0.936	−1	−0.929	−0.909
3	0.910	8.369	−0.931	−0.923	−1	−0.887	−0.813
4	0.817	12.577	−0.911	−0.911	−1	−0.868	−0.806
5	0.764	14.753	−0.934	−0.935	−1	−0.895	−0.821
6	0.635	17.522	−0.881	−0.876	−1	−0.805	−0.663
7	0.510	19.513	−0.900	−0.896	−1	−0.775	−0.623
8	0.385	22.573	−0.857	−0.870	−1	−0.691	−0.492
9	0.228	25.500	−0.818	−0.807	−1	−0.696	−0.543
10	0.130	28.657	−0.804	−0.765	−1	−0.653	−0.500

These results suggest that Q^2_{F1}, Q^2_{F2}, Q^2_{CCC}, and Q^2_{Rm} values depend on the relative distribution of the test set responses with respect to the training set responses. However, the estimate of the predictive ability of the model should depend only on the prediction error and not on the external set distribution. In other words, if the predicted ($\hat{y}$) and

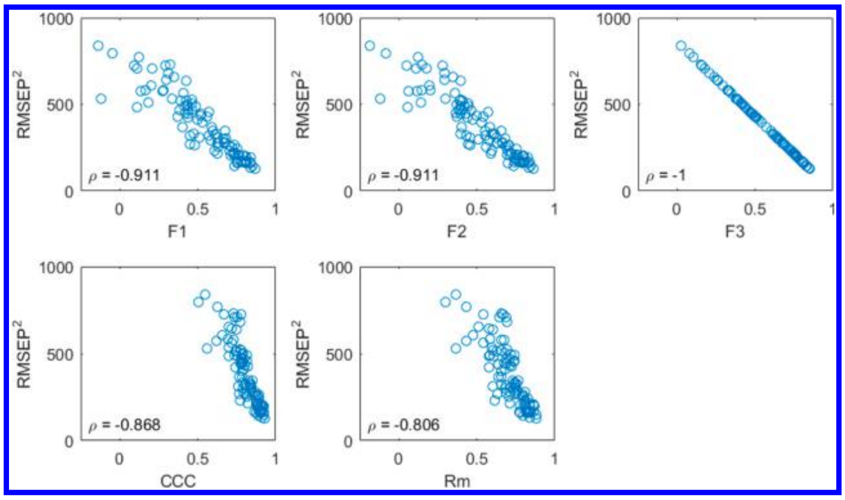


Figure 2. Q^2 values of the 100 repetitions of the five metrics vs squared *RMSEP* for *ID* = 4 (Table 3). The corresponding Pearson correlations (ρ) are also reported.

Table 4. Average Difference between the Q^2 Values Calculated on the Whole Test Set and the Average Value Calculated from 5 Test Set Partitions

ID	*RMSEC*	*RMSEP*	Q^2_{F1}	Q^2_{F2}	Q^2_{F3}	Q^2_{CCC}	Q^2_{Rm}
1	2.688	2.779	0.250	0.891	0.000	0.390	2.228
2	5.523	6.901	1.484	5.305	0.000	2.031	4.402
3	9.027	8.500	2.042	8.027	0.000	2.727	5.684
4	11.940	11.517	3.883	14.458	0.000	4.607	6.491
5	13.699	16.561	8.101	29.859	0.000	6.356	5.481
6	17.713	18.546	11.132	36.146	0.000	7.742	5.983
7	19.580	21.831	13.055	49.310	0.000	7.710	5.713
8	24.044	24.388	17.736	64.162	0.000	9.400	6.284
9	26.840	28.219	24.811	103.029	0.000	11.384	6.729
10	29.194	29.289	23.514	88.169	0.000	6.018	5.482

[a] Statistics were computed over 500 repetitions and for increasing *RMSEC* and *RMSEP* values.

experimental (y) responses are plotted, "the closer the data in such a plot lie to the line $y = \hat{y}$, the better the model is, as the predicted numerical values are very close to those measured by experiment."[7]

3.3. Principle 2: Correlation to RMSEP.

In order to evaluate the correlation of the analyzed Q^2 metrics to *RMSEP*, the following procedure was applied:

(a) randomly generate a set of 100 experimental responses for training objects with a uniform distribution in the range [0, 100]; generate the corresponding 100 calculated responses by adding a random error to the experimental training responses; the random error is uniformly distributed in the range $[-E_{TR}, +E_{TR}]$;

(b) randomly generate a set of 25 experimental responses for test objects with a uniform distribution in the range [0, 100]; generate the corresponding 25 predicted responses by adding a random error to the experimental test responses; the random error is uniformly distributed in the range $[-E_{TS}, +E_{TS}]$;

(c) repeat (b) for 100 iterations; in each iteration E_{TS} increases so that $E_{TS1} < E_{TS2} < ... < E_{TS100}$;

(d) for each iteration, calculate the corresponding *RMSEP* and Q^2 values; it derives that $RMSEP_1 < RMSEP_2 < ... < RMSEP_{100}$;

(e) calculate the Pearson correlation between the 100 Q^2 values and 100 squared *RMSEP* values;

(f) repeat the procedure (a–e) for 10 iterations, with increasing *RMSEC* and *RMSEP*; in each iteration, E_{TR} and E_{TS} increase in increments of 5 units from 5 to 50, so that $RMSEC_1 < RMSEC_2 < ... < RMSEC_{10}$ and $RMSEP_1 < RMSEP_2 < ... < RMSEP_{10}$.

The calculated correlation values are reported in Table 3.

As already explained in the theory section, the Q^2 values should be perfectly and inversely correlated to *RMSEP* ($\rho = -1$). In fact, as any Q^2 metric should account for the model predictivity, when the *RMSEP* decreases, the Q^2 should correspondingly increase, and *vice versa*. However, the results collected in Table 3 highlight that (1) Q^2_{F1}, Q^2_{F2}, Q^2_{CCC}, and Q^2_{Rm} are not perfectly anticorrelated to $RMSEP^2$ and (2) when the model quality decreases, Q^2_{F1}, Q^2_{F2}, Q^2_{CCC}, and Q^2_{Rm} tend to lose their correlation to *RMSEP*; that is, they account for information partially independent from *RMSEP*. This behavior is particularly pronounced for Q^2_{CCC} and Q^2_{Rm} metrics. On the contrary, the Q^2_{F3} metric always has a correlation equal to −1.

As an additional example, the scatter plots of 100 Q^2 values obtained for case *ID* 4 (Table 3) versus the squared *RMSEP* values are shown in Figure 2. In addition to what is highlighted above, it can be observed that Q^2_{CCC} and Q^2_{Rm} have very high values also for very large *RMSEP* values, further stressing their tendency to overestimate the real prediction potential.

Moreover, the pairwise correlations between Q^2 metrics were also calculated over the 10 simulated cases of Table 3. When the *RMSEP* values are low, all the metrics tend to converge and,

DOI: 10.1021/acs.jcim.6b00277
J. Chem. Inf. Model. 2016, 56, 1905−1913

Table 5. Statistics Calculated for each *RMSEP* Value over 500 Simulated Repetitions

ID	*mean RMSEC*	*mean* R^2	*RMSEP*	Q^2_{F1}	Q^2_{F2}	Q^2_{F3}	Q^2_{CCC}	Q^2_{Rm}
1	2.887	0.990	3.087	0.985 (±0.000)	0.985 (±0.000)	0.989 (±0.000)	0.992 (±0.000)	0.987 (±0.000)
2	6.634	0.947	6.168	0.953 (±0.001)	0.952 (±0.000)	0.954 (±0.002)	0.976 (±0.000)	0.951 (±0.000)
3	8.938	0.904	9.070	0.905 (±0.000)	0.905 (±0.000)	0.901 (±0.004)	0.953 (±0.000)	0.910 (±0.000)
4	10.389	0.869	10.746	0.856 (±0.001)	0.855 (±0.000)	0.86 (±0.005)	0.936 (±0.000)	0.894 (±0.000)
5	12.982	0.797	12.944	0.764 (±0.001)	0.763 (±0.000)	0.798 (±0.009)	0.889 (±0.000)	0.796 (±0.000)
6	13.284	0.787	14.129	0.790 (±0.001)	0.790 (±0.000)	0.759 (±0.009)	0.911 (±0.000)	0.836 (±0.000)
7	16.140	0.687	15.700	0.706 (±0.003)	0.702 (±0.000)	0.704 (±0.011)	0.874 (±0.000)	0.807 (±0.000)
8	17.312	0.639	17.073	0.636 (±0.007)	0.623 (±0.000)	0.649 (±0.012)	0.834 (±0.000)	0.734 (±0.000)
9	19.038	0.566	18.175	0.619 (±0.007)	0.597 (±0.000)	0.605 (±0.017)	0.836 (±0.000)	0.734 (±0.000)
10	19.667	0.532	19.102	0.579 (±0.003)	0.576 (±0.000)	0.559 (±0.019)	0.824 (±0.000)	0.731 (±0.000)

[a]Mean and standard deviation (in brackets) for each metric are reported, as well as average of *RMSEC* and R^2 over the repetitions.

as expected, their pairwise correlations are quite high; on the contrary, when *RMSEP* is large, their correlation behavior is different: for example, in the worst case (*ID* = 10, Table 4), Q^2_{F1} and Q^2_{F2} resulted to have a correlation equal to 0.967 while Q^2_{CCC} and Q^2_{Rm} a correlation equal to 0.940. On the contrary, in the same case, Q^2_{F3} resulted quite uncorrelated with the other metrics, with the highest correlations being equal to 0.653 with Q^2_{CCC} and 0.500 with Q^2_{Rm}.

3.4. Principle 3: Ergodic Property.

The ergodic property was evaluated by exploiting the same simulation procedure used to test the invariance to *RMSEP* (Principle 1). In this case, each test set (constituted by 25 objects) was divided into 5 subsets of 5 objects each. The Q^2 metrics values were then calculated on (1) the whole test set (25 objects) and (2) as the average of the values calculated on each subset. Then the absolute average differences of the two cases were calculated for each metric (Table 4).

As already demonstrated by Consonni et al.,[12] the ergodic property holds for Q^2_{F3} but not for Q^2_{F1} and Q^2_{F2}. A particular case is that of Q^2_{F2}, which cannot be calculated when the external objects are considered one at a time. The same happens for Q^2_{CCC} (eq 10), which results in 0/0. Finally, also Q^2_{Rm} cannot be calculated when only one object at a time is considered: in this case, in fact, the R^2 cannot be calculated, while $R^2{}_{0,XY}$ and $R'^2{}_{0,XY}$ are trivially equal to 1 (eq 12).

All of the metrics, except Q^2_{F3}, give different values when calculated on the test set as a whole or when it is divided into subsets (Table 4). The difference between the two ways of computing the metrics, in particular, tends to increase when increasing the *RMSEP*. The largest differences are observed for Q^2_{F1} and Q^2_{F2}, while those calculated for Q^2_{CCC} and Q^2_{Rm} are less pronounced but still remarkable. In conclusion, the only metric that fully satisfies the ergodic property is Q^2_{F3}.

3.5. Dependence on Training Set Variations.

With Q^2_{F1} and Q^2_{F3} being the only metrics influenced by the training set experimental responses (see eqs 9.1 and 9.3), their dependence on training set variations has been evaluated by means of a fixed test set, predicted with different training sets. In particular, the following procedure was applied:

(a) randomly generate a set of 100 test experimental responses with a uniform distribution in the range [0, 100]; generate the corresponding 500 predicted responses by adding a random error to the experimental test responses; the random error is uniformly distributed in the range $[-E_{TS}, +E_{TS}]$;

(b) randomly generate a set of 500 experimental responses for training objects with a uniform distribution in the range [0, 100]; generate the corresponding 100 calculated responses by adding a random error to the experimental training responses; the random error is uniformly distributed in the range $[-E_{TR}, +E_{TR}]$, with $E_{TR} = E_{TS}$.

(c) repeat (b) for 500 iterations; at each iteration the *RMSEP* is constant, because the predicted test set is always the same. Calculate the corresponding 500 Q^2 values;

(d) calculate the average and the standard deviation over the 500 Q^2 values;

(e) repeat the procedure (a–d) for 10 iterations, with increasing *RMSEP*, so that $RMSEP_1 < RMSEP_2 < \ldots < RMSEP_{10}$.

Standard deviations and average values calculated over the 500 repetitions of each trial are reported in Table 5. As expected, Q^2_{F2}, Q^2_{CCC}, and Q^2_{Rm} are invariant to training set changes, since their calculation is independent from training set responses. Therefore, these indices are associated with null standard deviations at each trial. However, the trend toward too optimistic predictive values is still confirmed for Q^2_{CCC} and Q^2_{Rm} for increasing *RMSEP*. On the contrary, Q^2_{F1} and Q^2_{F3} show variations due to predictions of the same test set by means of different training sets.

However, these variations are reasonably low in all considered simulations, as they mainly consist of differences at the third decimal place, or at the second decimal place for Q^2_{F3} in the worst modeling cases. In particular, as shown in Table 5, the standard deviations calculated over the 500 simulated repetitions increase when the model quality (R^2) decreases. Q^2_{F3} showed the maximum standard deviation (±0.019) in the most unfavorable modeling case (R^2 equal to 0.532).

Note that these results were obtained by using training and test sets with the same response distribution, as, by definition, the training and test set should be extracted from the same population. If extreme situations occur (such as test and training sets with significantly different distributions of experimental responses), higher variations of Q^2_{F1} and Q^2_{F3} can be achieved. However, this is not a real and practical case, since, as a rule of thumb, the test set should be representative of the entire training distribution. Moreover, in QSAR modeling, it is crucial to delimit the chemical space of prediction reliability (Applicability Domain,[26] AD), in order to prevent extrapolations. The appropriate use of the AD concept avoids the described (despite not likely) extreme situations, since the predictions of those test objects laying outside the training chemical space are considered as unreliable.

DOI: 10.1021/acs.jcim.6b00277
J. Chem. Inf. Model. 2016, 56, 1905−1913

Table 6. Summary of the Metric Compliance to the Introduced Principles P0−P3, where "yes" Indicates That the Principle Is Fulfilled and "no" That It Is Not Fulfilled

	P0		*P1*		*P2*		*P3*	
Q^2 *metric*	*scaling invariance*		*invariance to RMSEP*		*correlation to RMSEP*		*ergodic principle*	
Q^2_{F1}	yes		no	high variance	no	moderately uncorrelated	no	not fulfilled
Q^2_{F2}	yes		no	high variance	no	moderately uncorrelated	no	not fulfilled
Q^2_{F3}	yes		yes		yes		yes	
Q^2_{CCC}	yes		no	low variance	no	highly uncorrelated	no	not fulfilled
Q^2_{Rm}	no	not fully invariant	no	medium variance	no	highly uncorrelated	no	not fulfilled

Besides these considerations, it can be noted that the correlations of the Q^2_{F1} and Q^2_{F3} metrics with $RMSEP^2$ (Principle 2) are −0.993 and −1.000, respectively.

3.6. Thresholds of Q^2 Metrics. The Q^2 value is often strictly used to decide the acceptability or nonacceptability of a regression model by defining a threshold. Certainly, threshold values on Q^2 metrics are useful to identify suitable/unsuitable models. However, they are quite arbitrary and should be used to give only a suggestion about the usefulness of a model, taking into account the modeling purposes. It has been reasonably suggested that regression QSAR models are acceptable if they satisfy the following conditions, necessary but not sufficient:[2]

$$R^2 > 0.6 \wedge Q^2 > 0.5$$

However, the widespread use of *a priori* fixed thresholds on Q^2 can be, in some cases, questionable.[27,28] Since the aim of Q^2 metrics is to allow a comparison between model performances independently from the response measuring unit, the introduction of different thresholds on a metric-basis is counterintuitive and suggests that these metrics are not uniformly distributed in the upper-bounded range [−∞, 1]. Moreover, the acceptability of models on a Q^2-basis largely depends on the studied problem. For example, models of physicochemical properties are usually acceptable for Q^2 values larger than 0.75,[29−31] while complex biological properties are usually more difficult to model and Q^2 values greater than 0.55 can be regarded as acceptable,[32−34] at least at the beginning of the research.

In addition, a model with $R^2 = 0.9$ and $Q^2 = 0.6$ is likely to be affected by overfitting and/or other pathologies,[35] due to the very high difference (30%) between its fitting and prediction abilities. Instead of thresholds on the single metrics, it was suggested to evaluate thresholds on some differences between R^2 and Q^2 metrics.[36] In general, it would be preferable to have a difference between R^2 and Q^2 lower than a predefined threshold, such as, for example, $\Delta(R^2,Q^2) \leq 0.10$. However, larger differences do not necessarily indicate bad models, as in the case of some very flexible methods, like random forest or deep neural nets, which often show very high R^2 values, but also keep good predictive ability on the test set. Interested readers can find a critical analysis of the use of thresholds for Q^2 metrics in Roy et al.[27]

Finally, concerning the Q^2_{Rm} metric, other critical aspects were highlighted in Chirico and Gramatica.[28] Among them, it was outlined that in the case of experimental and predicted responses which perfectly lie on a line with origin at zero, Q^2_{Rm} is always equal to 1, independently from the slope; that is, the presence of a proportional bias is not accounted for. Later, to overcome this issue, an additional rule was proposed, defined as the threshold on the absolute difference between the two models calculated for the lines through the origin (eq 12): if $|\Delta r^2_m| \leq 0.2$, the model is acceptable.[25] This quantity appears as an *ad hoc* unsuitable solution; indeed, when experimental and predicted responses are perfectly aligned with origin at zero, the whole range of acceptable values $(-0.2 \leq \Delta r^2_m \leq +0.2)$ is characterized by $Q^2_{Rm} = 1$, independently of the bias.

3.7. Summary. All the calculations were repeated using a test set of 100 objects instead of 25, and all the derived considerations were confirmed. In Table 6, a summary of the metric behavior is given for all the considered principles.

4. CONCLUSIONS

This study dealt with the evaluation of five Q^2 metrics (i.e., Q^2_{F1}, Q^2_{F2}, Q^2_{F3}, Q^2_{CCC}, and Q^2_{Rm}) designed to estimate the predictive ability of regression QSAR models. Four fundamental mathematical principles, which should be respected by any Q^2 metric, were introduced: (1) invariance to response scaling, (2) invariance to *RMSEP*, (3) correlation to *RMSEP*, and (4) ergodic principle. By means of simulated case studies, the metric compliance to the presented principles was analyzed and critically discussed.

The only Q^2 metric that satisfied all of the theoretical principles was Q^2_{F3}, while the other metrics showed several drawbacks. In particular, the results can be summarized as follows:

1. Q^2_{Rm} is not invariant to the response scaling;
2. Q^2_{F1} and Q^2_{F2} showed a remarkable variability on test sets having the same *RMSEP*:
3. Q^2_{CCC} systematically overestimates the model prediction ability;
4. Q^2_{Rm} overestimates the model prediction ability, especially in correspondence of high *RMSEP* values;
5. Q^2_{Rm} and Q^2_{CCC} are quite uncorrelated with *RMSEP*, especially in correspondence of decreasing model quality.

Given these considerations and under the assumptions of this study, the usage of Q^2_{F3} is strongly recommended, due to its demonstrated mathematical properties. For the same reasons, it is here suggested not to use Q^2_{F1}, Q^2_{F2}, Q^2_{CCC}, and Q^2_{Rm} metrics because they appear unreliable in evaluating the prediction ability of regression models.

Actually, in the common practice, good indications on the acceptability of a new model can be obtained by comparing its prediction ability (Q^2), along with its simplicity and interpretability, with other models proposed in the literature, if available. Moreover, also the "distance" of each predicted object from the training set, i.e. the applicability domain, should be evaluated to avoid doubtful extrapolations.

■ AUTHOR INFORMATION

Corresponding Author

*Tel: +39-0264482820. E-mail: roberto.todeschini@unimib.it.
URL: http://michem.disat.unimib.it/chm/.

Notes

The authors declare no competing financial interest.

REFERENCES

(1) Golbraikh, A.; Shi, L. M.; Xiao, Z.; 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.

(2) Tropsha, A.; Gramatica, P.; Gombar, V. K. The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models. *QSAR Comb. Sci.* **2003**, *22*, 69−77.

(3) Schmuker, M.; Givehchi, A.; Schneider, G. Impact Of Different Software Implementations on the Performance of the Maxmin Method for Diverse Subset Selection. *Mol. Diversity* **2004**, *8*, 421−425.

(4) Fourches, D.; Barnes, J. C.; Day, N. C.; Bradley, P.; Reed, J. Z.; Tropsha, A. Cheminformatics Analysis of Assertions Mined from Literature that Describe Drug-Induced Liver Injury in Different Species. *Chem. Res. Toxicol.* **2010**, *23*, 171−183.

(5) Hsieh, J.-H.; Yin, S.; Liu, S.; Sedyckh, A.; Dokholyan, N. V.; Tropsha, A. Combined Application of Cheminformatics- and Physical Force Field-Based Scoring Functions Improves Binding Affinity Prediction for CSAR Data Sets. *J. Chem. Inf. Model.* **2011**, *51*, 2027−2035.

(6) Colmenarejo, G. In Silico Prediction of Drug-Binding Strengths to Human Serum Albumin. *Med. Res. Rev.* **2003**, *23*, 275−301.

(7) Alexander, D. L. J.; Tropsha, A.; Winkler, D. A. Beware Of R^2: Simple, Unambiguous Assessment of the Prediction Accuracy of QSAR and QSPR Models. *J. Chem. Inf. Model.* **2015**, *55*, 1316−1322.

(8) Grisoni, F.; Consonni, V.; Vighi, M.; Villa, S.; Todeschini, R. Expert QSAR System for Predicting the Bioconcentration Factor under the REACH Regulation. *Environ. Res.* **2016**, *148*, 507.

(9) Todeschini, R.; Consonni, V. *Molecular Descriptors for Chemoinformatics (2 volumes)*; WILEY-VCH: Weinheim (Germany), 2009.

(10) OECD Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models. *OECD Series on Testing and Assessment*; 2014, 69.

(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) 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.

(14) Mitra, I.; Roy, P. P.; Kar, S.; Ojha, P. K.; Roy, K. On Further Application of R^{2m} as a Metric For Validation of QSAR Models. *J. Chemom.* **2010**, *24*, 22−33.

(15) Ojha, P. K.; Mitra, I.; Das, R. N.; Roy, K. Further Exploring R^{2m} Metrics for Validation of QSPR Models. *Chemom. Intell. Lab. Syst.* **2011**, *107*, 194−205.

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

(17) Gramatica, P.; Sangion, A. A Historical Excursus on the Statistical Validation Parameters for QSAR Models: A Clarification Concerning Metrics And Terminology. *J. Chem. Inf. Model.* **2016**, *56*, 1127−1131.

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

(19) Lin, L. I. Assay Validation Using The Concordance Correlation Coefficient. *Biometrics* **1992**, *48*, 599−604.

(20) Golbraikh, A.; Tropsha, A. Predictive QSAR Modeling Based on Diversity Sampling of Experimental Datasets for the Training and Test Set Selection. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 357−369.

(21) Martin, T. M.; Harten, P.; Young, D. M.; Muratov, E. N.; Golbraikh, A.; Zhu, H.; Tropsha, A. Does Rational Selection of Training and Test Sets Improve the Outcome of QSAR Modeling? *J. Chem. Inf. Model.* **2012**, *52*, 2570−2578.

(22) Roy, K.; Chakraborty, P.; Mitra, I.; Ojha, P. K.; Kar, S.; Das, R. N. Some Case Studies on Application of "R^{2m}" Metrics for Judging Quality of Quantitative Structure−Activity Relationship Predictions: Emphasis on Scaling of Response Data. *J. Comput. Chem.* **2013**, *34*, 1071−1082.

(23) Roy, K.; Mitra, I.; Ojha, P. K.; Kar, S.; Das, R. N.; Kabir, H. Introduction Of R^{2m} (Rank) Metric Incorporating Rank-Order Predictions as an Additional Tool for Validation of QSAR/QSPR Models. *Chemom. Intell. Lab. Syst.* **2012**, *118*, 200−210.

(24) Atkinson, G.; Nevill, A. Comment on The Use of Concordance Correlation to Assess the Agreement Between Twovariables. *Biometrics* **1997**, *53*, 775−777.

(25) Roy, K.; Das, R. N.; Ambure, P.; Aher, R. B. Be Aware of Error Measures. Further Studies on Validation of Predictive QSAR Models. *Chemom. Intell. Lab. Syst.* **2016**, *152*, 18−33.

(26) Sahigara, F.; Mansouri, K.; Ballabio, D.; Mauri, A.; Consonni, V.; Todeschini, R. Comparison of Different Approaches to Define The Applicability Domain of QSAR Models. *Molecules* **2012**, *17*, 4791−4810.

(27) Roy, K.; Mitra, I.; Kar, S.; Ojha, P. K.; Das, R. N.; Kabir, H. Comparative Studies on Some Metrics for External Validation of QSPR Models. *J. Chem. Inf. Model.* **2012**, *52*, 396−408.

(28) Chirico, N.; Gramatica, P. Real External Predictivity of QSAR Models. Part 2. New Intercomparable Thresholds for Different Validation Criteria And the Need for Scatter Plot Inspection. *J. Chem. Inf. Model.* **2012**, *52*, 2044−2058.

(29) De Lima Ribeiro, F. A.; Ferreira, M. M. C. QSPR Models of Boiling Point, Octanol−Water Partition Coefficient And Retention Time Index of Polycyclic Aromatic Hydrocarbons. *J. Mol. Struct.: THEOCHEM* **2003**, *663*, 109−126.

(30) Yao, X. J.; Panaye, A.; Doucet, J. P.; Zhang, R. S.; Chen, H. F.; Liu, M. C.; Hu, Z. D.; Fan, B. T. Comparative Study of QSAR/QSPR Correlations Using Support Vector Machines, Radial Basis Function Neural Networks, And Multiple Linear Regression. *J. Chem. Inf. Comp. Sci.* **2004**, *44*, 1257−1266.

(31) Grisoni, F.; Cassotti, M.; Todeschini, R. Reshaped Sequential Replacement Algorithm for Variable Selection in QSPR Modelling: Comparison with Other Benchmark Methods. *J. Chemom.* **2014**, *28*, 249−259.

(32) Cronin, M. T.; Aptula, A. O.; Duffy, J. C.; Netzeva, T. I.; Rowe, P. H.; Valkova, I. V.; Schultz, T. W. Comparative Assessment of Methods to Develop QSARs for the Prediction of the Toxicity of Phenols to *Tetrahymena Pyriformis*. *Chemosphere* **2002**, *49*, 1201−1221.

(33) Votano, J. R.; Parham, M.; Hall, L. M.; Hall, L. H.; Kier, L. B.; Oloff, S.; Tropsha, A. QSAR Modeling of Human Serum Protein Binding with Several Modeling Techniques Utilizing Structure-Information Representation. *J. Med. Chem.* **2006**, *49*, 7169−7181.

(34) Tetko, I. V.; Sushko, I.; Pandey, A. K.; Zhu, H.; Tropsha, A.; Papa, E.; Õberg, T.; Todeschini, R.; Fourches, D.; Varnek, A. Critical Assessment of QSAR Models of Environmental Toxicity Against *Tetrahymena Pyriformis*: Focusing on Applicability Domain and Overfitting by Variable Selection. *J. Chem. Inf. Model.* **2008**, *48*, 1733−1746.

(35) Todeschini, R.; Consonni, V.; Mauri, A.; Pavan, M. Detecting "Bad" Regression Models: Multicriteria Fitness Functions in Regression Analysis. *Anal. Chim. Acta* **2004**, *515*, 199−208.

(36) Sutter, J. M.; Peterson, T. A.; Jurs, P. C. Prediction of Gas Chromatographic Retention Indices of Alkylbenzene. *Anal. Chim. Acta* **1997**, *342*, 113−122.

DOI: 10.1021/acs.jcim.6b00277