Chemometrics and Intelligent Laboratory Systems 152 (2016) 18–33

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

Be aware of error measures. Further studies on validation of predictive QSAR models☆

Kunal Roy *, Rudra Narayan Das [1], Pravin Ambure [1], Rahul B. Aher [1]

Drug Theoretics and Cheminformatics Laboratory, Department of Pharmaceutical Technology, Jadavpur University, Kolkata 700 032, India

ARTICLE INFO

Article history:
Received 18 November 2015
Received in revised form 8 January 2016
Accepted 10 January 2016
Available online 18 January 2016

Keywords:
QSAR
Validation
Error measures
Dispersion
MAE

ABSTRACT

Validation is the most crucial concept for development and application of quantitative structure–activity relationship (QSAR) models. The validation process confirms the reliability of the developed QSAR models along with the acceptability of each step during model development such as assessing the quality of input data, dataset diversity, predictability on an external set, domain of applicability and mechanistic interpretability. External validation or validation using an independent test set is usually considered as the gold standard in evaluating the quality of predictions from a QSAR model. The external predictivity of QSAR models is commonly described by employing various validation metrics, which can be broadly categorized into two major classes, viz., R^2 based metrics namely R^2_{test}, $Q^2_{(ext\_F1)}$, and $Q^2_{(ext\_F2)}$, and purely error based measures like predicted residual sum of squares (PRESS), root mean square error (RMSE), and mean absolute error (MAE). The problem associated with the error based measures is the absence of any well-defined threshold for determining the quality of predictions making the R^2 based metrics more suitable for use due to easy comprehension. However, in this paper, we show the problems associated with the R^2 based validation metrics commonly used in QSAR studies, since their values are highly dependent on the range of the response values of the test set compounds and their distribution pattern around the training/test set mean. We also propose a guideline for determining the quality of predictions based on MAE and its standard deviation computed from the test set predictions after omitting 5% high residual data points in order to obviate the influence of any rarely occurring high prediction errors that may significantly obscure the quality of predictions for the whole test set. In this manner, we try to evaluate the prediction performance of a model on most (95%) of the data points present in the external set. An online tool (XternalValidationPlus) for computing the suggested MAE based criteria (along with other conventional metrics) for external validation has been made available at http://dtclab.webs.com/software-tools and http://teqip.jdvu.ac.in/QSAR_Tools/. The MAE based criteria suggested here along with other commonly used validation metrics may be applied to evaluate predictive performance of QSAR models with a greater degree of confidence.

1. Introduction

Quantitative structure–activity relationship (QSAR) modeling is utilized in rational drug design, environmental risk assessment and fate modeling, toxicity and property prediction of chemicals and pharmaceuticals. A QSAR model represents a mathematical relationship for a set of molecules (*training set*) with a known response (*activity/toxicity/property*) obtained from application of various chemometric techniques (statistical tools). The actual relationship is built between the structural features of a molecule expressed in quantitative terms (*descriptors/independent variables*) that are derived computationally (or experimentally in some cases) and the dependent variable or the response, which should always be experimentally derived. QSAR is a time- as well as cost-effective technique that supports the 3Rs (*replacement, refinement and reduction in animals in research*) paradigm [1]. The chemical and toxicological regulatory agencies worldwide have been employing QSAR models for decision-making frameworks in risk and safety assessments for a number of years [2].

Validation is the most crucial concept for development and application of any QSAR model. The validation process confirms the reliability of the developed QSAR model along with the acceptability of each step during model development such as assessing the quality of input data, dataset diversity, predictability on external set, domain of applicability and mechanistic interpretability. For regulatory acceptance of QSAR models, five guidelines are agreed by the Organization for Economic Co-operation and Development (OECD) [3], and these cover the following criteria: (i) a defined endpoint, (ii) an unambiguous algorithm, (iii) the domain of applicability, (iv) appropriate measures of goodness of fit, robustness and predictivity of the developed model, and (v) a

☆ Presented in part at the 23rd Conference on Current Trends in Computational Chemistry (23rd CCTCC) held in Jackson, Mississippi, USA, on November 13–14, 2015.

* Corresponding author. Tel.: +91 98315 94140; fax: +91 33 2837 1078.
E-mail addresses: kunalroy_in@yahoo.com, kroy@pharma.jdvu.ac.in (K. Roy).
URL: http://sites.google.com/site/kunalroyindia/ (K. Roy).
[1] Tel.: +91 98,315 94,140; fax: +91 33 2837 1078.

http://dx.doi.org/10.1016/j.chemolab.2016.01.008
0169-7439/

mechanistic interpretation, if possible. These guidelines are now referred to as OECD Principles for the validation of QSARs. As a result, different groups of researchers have shown their keen interest towards the development of more appropriate validation metrics for precise and predictive QSAR model development [3–8].

Recently, Alexander et al. [9] have suggested some shortcomings in the model fit criteria previously suggested by Golbraikh and Tropsha, [5] and further proposed that only two simple criteria might be sufficient for judging model usefulness: high R^2 (*correlation coefficient*) and low root mean square error (RMSE) of test set predictions. In the present paper, we have tried to relook the problem in a greater detail. We have highlighted some problems associated with the conventional R^2 based validation metrics and suggested a set of criteria based on model errors for unbiased judgment of the quality of model predictions. Although the problems associated with the R^2 based formalism were identified long back [10], QSAR practitioners usually rely on these metrics for evaluating the predictive potential of models. Therefore, we aim to provide here the readers with an overview of the shortcomings of the conventional R^2 based metrics as applied in QSAR studies and also encourage easy interpretation of the quality of predictions from the error based judgment.

2. Problems with the conventional metrics

Validation of QSAR models is a crucial issue for judging their ability of prediction for the chemicals not employed during model development. In consonance with the OECD guidelines regarding the model fitness, robustness as well as predictivity, a number of statistical metrics are used by different research groups in this field. In this article, we shall restrict our discussion to the validation aspect involving test set compounds only, i.e., metrics characterizing external validation of a model.

A commonly used regression based measure is determination coefficient (R^2) between the observed and predicted response values of the test set compounds. This metric may be computed based on the following expression [11].

$$R^2 = \frac{\left[\sum\left\{\left(Y_{obs}-\overline{Y_{obs}}\right)\times\left(Y_{pred}-\overline{Y_{pred}}\right)\right\}\right]^2}{\sum\left(Y_{obs}-\overline{Y_{obs}}\right)^2\times\sum\left(Y_{pred}-\overline{Y_{pred}}\right)^2} \quad (1)$$

In Eq. (1), Y_{obs} and Y_{pred} correspond to the observed (i.e., experimental) and predicted response values respectively of the test set compounds. Instead of providing a true picture of the prediction errors encountered, the R^2 metric as defined in Eq. 1 attempts to provide a relative pattern of changes in the values of the observed response with respect to the predicted ones. As a result, this metric can furnish acceptable values for a constant magnitude of errors for all the samples even if it is very high. A way out to overcome this problem may be the use of the regression through origin (RTO) approach where the best fitted line is deliberately forced through origin ($Y_{obs}=0, Y_{pred}=0$) in order to penalize the R^2 value obtained from the corresponding normal regression analysis in case of large prediction errors [11]. Based on the judgment of RTO derived method, researchers in this field have formulated model validation criteria such as Golbraikh and Tropsha's criteria [6] as well as different r_m^2 metrics [7,12,13].

However, the RTO approach is able to identify prediction errors of a model as long as the data are devoid of any 'systematic error' and/or model bias. Systematic error is usually characterized by bias in model predictions. However, such errors in analytical experiments are avoidable and mostly represent those arising from operational perspective of the analyst, instrumental adjustments, reagent based defects, as well as improper method based flaws giving inaccurate results [14]. Dearden et al. [15] have identified such errors in models due to improper selection of model variables. A biased model prediction may be characterized by all error values of same sign, i.e., all (or disproportionately high fractions) being positives or all (or disproportionately high fractions) being negatives. The determination coefficient R^2 and its origin based counterpart, i.e., R_0^2 are applicable only if the predicted data are devoid of such existing 'systematic error' feature, otherwise it might give a wrong assessment of the model predictivity. In case of the presence of any systematic error or model bias for a particular test set, attempt should be made to change the model to remove the systematic error as such test set is not suitable for predictions from the developed model in any validation experiment. This is something similar to adjusting instrumental error before doing an instrumental analysis and such error has nothing to do with the quality of determinations (predictions in our case). Some methods for the identification of systematic error include residual plot analysis [15], implementation of Kriging models [16], comparison analysis involving average error and average absolute error measure [17].

The external predictivity of QSAR models is commonly described by employing various validation metrics, and these can be broadly categorized into two major classes, viz., R^2 based metrics namely $Q^2_{ext(F1)}$ and $Q^2_{ext(F2)}$ [18] and error based measures like predicted residual sum of squares (PRESS), standard error of estimate (SEE), root mean square error (RMSE), and mean absolute error (MAE) [19]. An alternative general formula for R^2 has been furnished in Eq. 2 which is most commonly used for computation of different R^2 based validation metrics (or Q^2_{ext}).

$$R^2 = 1-\frac{\sum\left(Y_{obs}-Y_{pred}\right)^2}{\sum\left(Y_{obs}-\overline{Y}\right)^2} \quad (2)$$

In Eq. 2, the experimental and predicted response values of a chemical have been designated using Y_{obs} and Y_{pred} respectively, while $\overline{Y}$ represents the mean response value of the training set or the test set compounds, depending upon the metric used (for example, $Q^2_{ext(F1)}$ and $Q^2_{ext(F2)}$). $Q^2_{ext(F1)}$ uses the training set mean value while it is the test set mean in the case of $Q^2_{ext(F2)}$.The numerator of the fraction shown in Eq. 2 is a measure of prediction error and the R^2 metric measures the model performance (in terms of prediction errors) in comparison to the performance of a "no model" situation (that is the mean of the response values of the training or test set compounds considered as the reference). A model will be of no use if its prediction performance is not better than, at least, the performance of the mean (i.e., "no model"). Thus, a model may be considered acceptable when the values of these R^2 based metrics ($Q^2_{ext(F1)}$ and $Q^2_{ext(F2)}$) are at least more than 0.5; the closer are the values to unity, the greater is the confidence in prediction precisions. Although this formalism seems logical, the results from the comparison with the performance of the mean can sometimes be misleading since it is greatly influenced by the range of the corresponding training/test set data, the average value of which is used in Eq. 2. Another important aspect for the over- and under-estimation of prediction errors by the Q^2_{ext} metrics is the distribution of the response data around mean. In the case of the test set mean based assessment, i.e., $Q^2_{ext(F2)}$, if most of the response data points remain in close neighborhood of the mean value leaving only a low fraction away from it, the mean can perform well as an estimate of the individual responses and the value of the metric can be low in spite of the presence of low amount of prediction errors. Thus, the judgment provided by the Q^2_{ext} metrics is not only dependent on model based predictions but also on other factors like range as well as distribution of the response data around mean, and therefore such metrics cannot be identified as sufficient measures for external validation. We may mention here that the expression of R^2 for external validation as suggested by Alexander et al. [9] actually corresponds to $Q^2_{ext(F2)}$, though they did not mention about it explicitly in their paper.

While the Q^2_{ext} metrics may give misleading results regarding the quality of predictions, prediction error based metrics like PRESS, SEE, MAE, and RMSE [19] give more straight-forward results. Now, the main problem while dealing with such metrics is the absence of a suitable threshold value unlike the Q^2_{ext} metrics. It is to be noted that

like other statistical measures, these error based metrics also condense a significant amount of error data into a single value, and obviously they can therefore portray only one projection of the prediction error values. Hence, it might be difficult to judge the amount of prediction errors encountered from competing models. In other words, one cannot simply detect the presence of small number of high prediction errors or a significant number of moderate prediction errors from the value of a single error based metric. Therefore, determination of an error based metric may not be sufficient, and further introspection is needed to reflect the distribution of errors. An expanded version of this section along with explanatory illustrations is available in the Supplementary material section.

3. Legitimate judgment of model predictivity in terms of prediction errors

The prime objective of external validation of a QSAR model is a simple judgment of the errors encountered during prediction of a test set within the model domain. By the term domain here, we mean the chemical and response based domains of the training set. Now, we can see that although widely employed, the Q^2_{ext} based metrics are unable to provide a true reflection of prediction errors because of the influence of the data range as well as distribution of the response values around the mean response value of the training/test set compounds. However, there is no doubt that in case of an appropriate distribution of the responses around the (training/test) mean and a test set response range which is not too wide or too narrow, the Q^2_{ext} metrics can be useful determinants of model quality. For development of a QSAR model, usually a suitable range for the training set response which is not less than 3–4 log units is considered, but at present there is no such specific requirement of the range prescribed for the test set response data. However, it appears that for the test set which is used for the validation of a QSAR model, the range of the response values of the test set and distribution of the responses around the mean are also very important.

The determination of the error based metrics gives more direct information about the prediction errors since they do not compare the prediction errors with other aspects like performance of the mean. Again, a single metric condensing the error values of the whole test set may not reflect non-uniform distribution of prediction errors among the data points; thus, a study on the distribution of errors seems reasonable. A determination of standard deviation of prediction errors may indicate the presence (or absence) of a considerable number of outlier predictions in the test set which would otherwise remain unnoticed for a relatively large amount of data points where the mean error can still have a low value. Now, the prediction errors in a test set may be present either in a small number with high error values or at a significant number but with moderate error values. Let us consider a test set prediction where most of the compounds are well predicted except only a few of them being badly predicted. We may have another test set where the predictive performance of the model on the test set compounds is moderate for most of the compounds. Now, using an error based metric it may be shown that the metric value is same or close to each other in both the cases but it does not give any idea on the distribution of the errors. It is quite obvious that the predictive performance of the former case (in the given example) is better where most of the predictions are good. Hence, it is logical to check how the model performs for most (say 95%) of the compounds in the test set, as there may be a limited number of outlier predictions which may significantly affect the values of the error based metrics computed from all test set compounds. Therefore, one should monitor two aspects of the prediction errors namely judgment of the dispersion of errors, and secondly comparison of the values of the error based metrics after removing a definite small (say 5%) fraction of chemicals possessing high residual values for that test set. Fig. 1 presents a hypothetical case where most of the compounds in a test set are well predicted except a few bad predictions, and Fig. 2 shows a case where most of the predictions can be considered moderate. In both cases of Fig. 1 and Fig. 2, the MAE value has been kept constant and it can be observed that the goodness of predictions is explainable when 5% of high residuals data are omitted.

The two most commonly used error based metrics in QSAR literature are RMSE and MAE, the formulae of which have been depicted in Eqs. 3 and 4, respectively. It is evident that both of the metrics provide measures of the actual prediction errors with respect to the total number of observations. However, the information provided by RMSE has been argued to be more complicated than MAE [20].

$$RMSE = \sqrt{\frac{1}{n} \times \sum \left(Y_{obs} - Y_{pred}\right)^2} \tag{3}$$

$$MAE = \frac{1}{n} \times \sum \left|Y_{obs} - Y_{pred}\right| \tag{4}$$

Because of the involvement of squared term of the prediction errors in the expression of RMSE, the variance of errors may be influenced for a set of data. That is, squaring the higher prediction error values will have more weight than the lower errors in the formalism of RMSE while MAE provides an equal weight to all the errors. Thus, MAE is considered to be a simpler and more straight-forward determinant of prediction errors [20]. Willmott and Matsuura [21] have shown RMSE to be a function of three parameters namely magnitude of the prediction errors in the

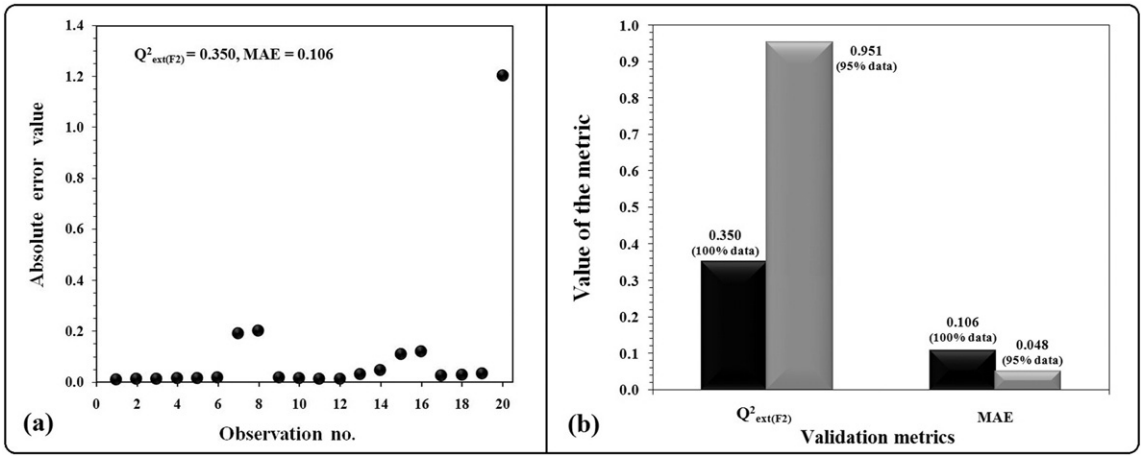

Fig. 1. The $Q^2_{ext(F2)}$ metric portraying an unacceptable model prediction quality due to the presence of a relatively small amount of high residual observations: A) Plot of the MAE values showing most of them close to zero except a small number of influential data points, B) comparative analysis showing significant change in the values of the $Q^2_{ext(F2)}$ and MAE metric when 5% of high residual observations are removed.

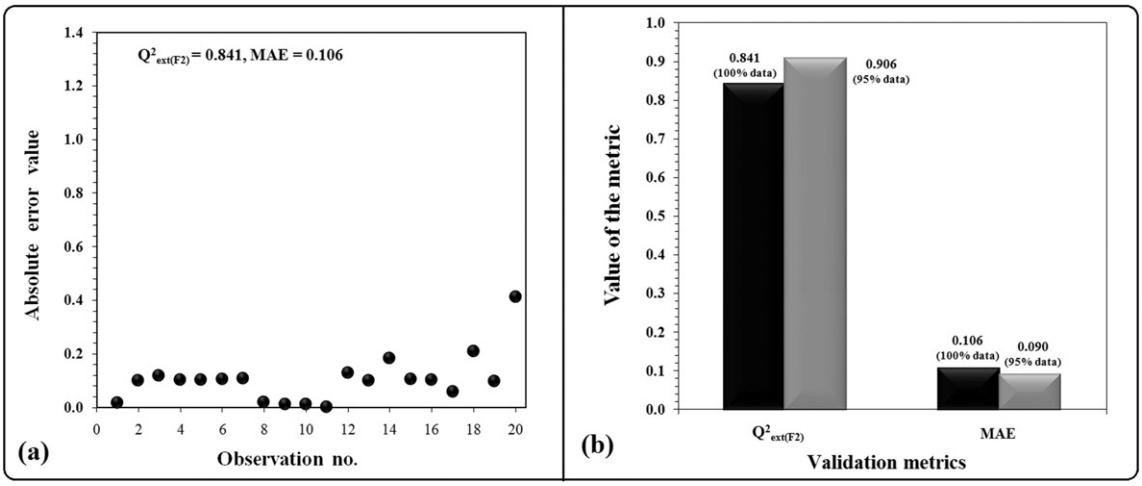

Fig. 2. The $Q^2_{ext(F2)}$ metric portraying an acceptable model prediction quality with most of the observations possessing moderate predicted residual values: A) Plot of the MAE values showing most of them with moderate values, B) comparative analysis depicting less significant change in the values of the $Q^2_{ext(F2)}$ and MAE metrics when 5% of high residual observations are removed.

squared form, the magnitude of the average error, and squared root of the number of samples while MAE is devoid of such complex parameterization. Therefore, determination of the central tendency that is average prediction error along with the error dispersion is more straightforward and direct in the case of MAE. Hence, we have chosen MAE (also known as average absolute error (AAE)) to be a better index of errors in the context of predictive modeling studies. Along with the determination of MAE, the computation of the standard deviation of the absolute error values is also encouraged. Now, there is another concern regarding the threshold value of any error based metric. Since, the number of data points may be variable for different test sets, and no comparison is made here with respect to the mean based performance of a model prediction, it will be incorrect to assign a specific value to be the limiting one for different datasets. Instead, a given percentage of errors with respect to the response range can be a suitable indicator of model predictivity. Glick [22] reported the usefulness of the concept of a given error percentage with respect to clinical data in the late seventies. Now, it is to be understood that the magnitude of error is actually a relative feature and can vary greatly depending upon the purpose of the analysis employed. Upon a closer inspection, we will be able to comprehend that the allowable magnitude of error varies from case to case depending on the range of the data employed. Considering the perspective of QSAR analysis, composition of the training set presents the response domain of a model and remains unchanged while the test set composition may vary. This is so because one can also desire to judge the performance of a model by employing separate test sets with varying compositions. Thus, consideration of the range of the training set provides a reasonable basis for the determination of threshold value for the error measure (MAE) for the external validation. Here, we propose the following criteria to be checked while assessing external predictivity of a QSAR model. Note that we consider the response (Y) values in logarithmic units only, as customary in most QSAR analyses. If the response values are in raw data form (not in log units), then they must be converted to log units before application of the following criteria.

i) Good predictions:
From a general notion, an error of 10% of the training set range should be acceptable while an error value more than 20% of the training set range should be a very high error. Thus, the criteria for good predictions could be the following:

$$\text{MAE} \le 0.1 \times \text{training set range AND MAE} + 3 \times \sigma \le 0.2 \times \text{training set range}.$$

Here, the σ value denotes the standard deviation of the absolute error values for the test set data. Considering a normal distribution pattern, mean $\pm$ 3σ covers 99.7% of the data points.

ii) Bad predictions:
A value of MAE more than 15% of the training set range should be high while an error more than 25% of the training set range is considered very high. Hence, the predictions could be considered bad when:

$$\text{MAE} > 0.15 \times \text{training set range OR MAE} + 3 \times \sigma > 0.25 \times \text{training set range}.$$

The predictions which do not fall under either of the above two conditions may be considered as of moderate quality. The above criteria should be applied for judging the quality of test set predictions when the number of data points is at least 10 (statistical reliability) and there is no systematic error in model predictions (statistical applicability).

Again, considering that most of the statistical tests are usually performed at the probability level of 5% [23], we propose that both the abovementioned MAE based criteria might be determined after removing 5% test set chemicals with high residual values in order to obviate the possibility of any outlier predictions. The judgment based on the 95% data will be very helpful for test sets containing small number of badly predicted data points where the MAE based criteria for the 100% data can penalize the model predictivity. In such cases, comparison of the proposed MAE based criteria for the 100% and 95% data can allow a quick identification of the occurrence of small number of high residual data points, if any. It is to be noted that the descriptors used in a predictive modeling analysis attempt to capture the chemical information of the chemicals used for developing the equation. Now, unsatisfactory predictive performance for a small number of the observations by the QSAR model observed during the external validation test might actually portray the deficit in chemical information captured by the model or indicate presence of experimental errors in the observed data. Since QSAR models are applicable only within a defined domain of the data used for developing them, such small number of influential observations (which shifts the value of an error based metric significantly from that obtained without considering these observations) are usually identified as 'outliers'. Hence, we strongly believe that there is a need for contemplation on the possible occurrence of influential observations while judging the predictivity by using an external set which might actually influence

the final error based metric computed. In that sense, a model providing reasonable predictive performance for most of the chemicals can be considered to be a good model. Computation of the metric values after the removal of 5% high prediction error data points allows judging the quality of the model for most of the compounds. The above-mentioned criteria along with the classical validation metrics can be computed using an online tool 'XternalValidationPlus' developed by the present authors' group (available at http://dtclab.webs.com/software-tools and http://teqip.jdvu.ac.in/QSAR_Tools/). The basic workflow of the algorithm implemented in 'XternalValidationPlus' is furnished in Fig. 3.

4. Results and discussion

In this present communication, we have attempted to judge the performance of various conventional validation metrics along with our proposed MAE based criteria by employing test sets with varying compositions of the response values in terms of range as well as their distribution around the mean of the training/test set. We have chosen, for this purpose, three published datasets previously used for developing predictive in silico models towards physicochemical property, activity, and toxicity endpoints. The first dataset represents aqueous solubility of drug like molecules and agrochemicals [24], the second set is on acetylcholinesterase inhibitory activity of functionalized organic chemicals [25], and the third one is a dataset comprising cytotoxicity of ionic liquids to the rat [26]. In all the three cases, we have used the predicted response values of the test sets using the reported QSPR/QSAR/QSTR models. Now, in order to construct test sets of varying composition, we have developed various subsets of the whole test set using a window based approach. After taking the experimental and predicted values for the whole test set (from the published studies), we have sorted the data based on the observed response, and then test subsets were selected by taking compounds from the top, median, and bottom zones for nine different logarithmic ranges viz., 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, and 4.5. Any test subset comprising fewer than or equal to 10 compounds was not considered. The same formalism was implemented for all the three mentioned datasets. Fig. 4 presents an overview of the subset division from the whole test set. Since the test subsets are picked up from different response zones of the test sets, the distribution of compounds around the training/test set mean varies along with the test set range. Now, considering the use of the same model for deriving predictions, it is interesting to note that the metrics are supposed to judge the quality of the same model on different test subsets of varying response compositions. Furthermore, the values of different validation metrics were also computed for the whole test set enabling a comparison analysis of the subsets with the whole test set range as employed in the published studies.

In order to validate the consistency of our proposed MAE parameter based criteria along with other conventional metrics, we have performed a Y-randomization analysis using the final descriptors of

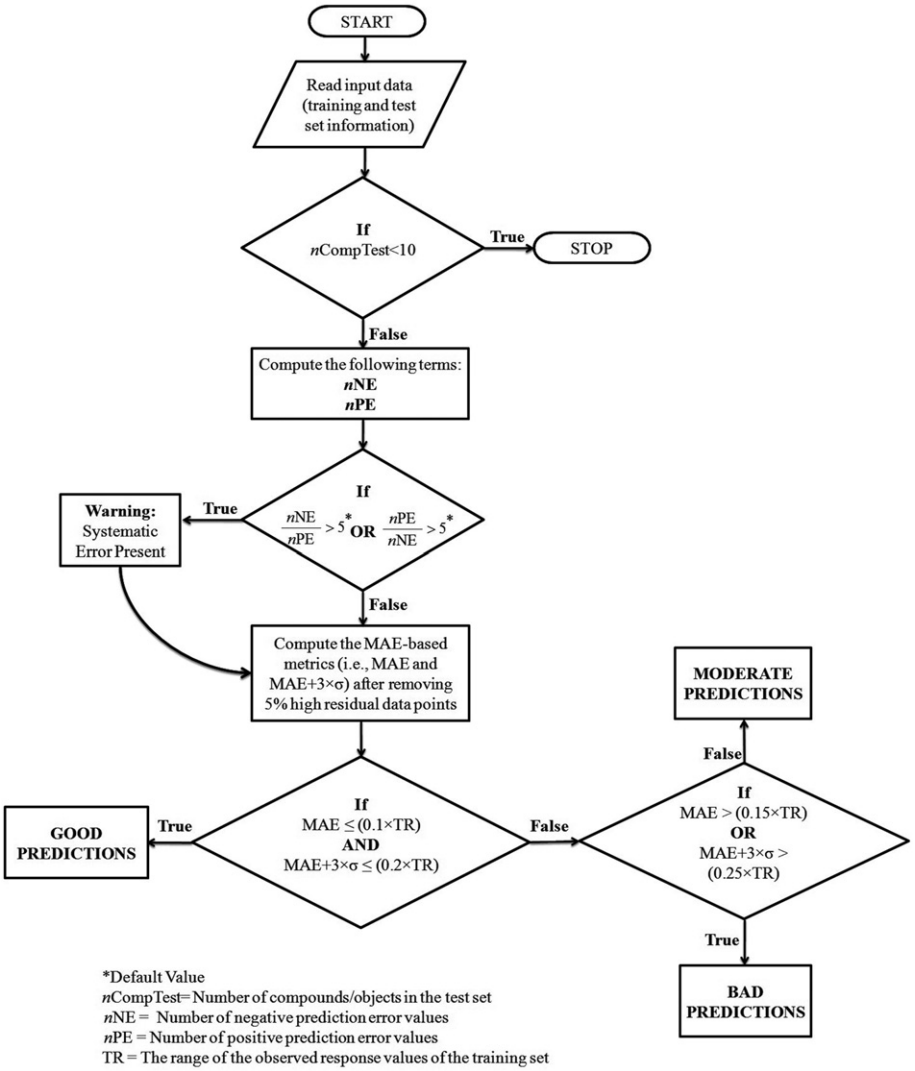

Fig. 3. The basic algorithmic workflow as used in the tool XternalValidationPlus.

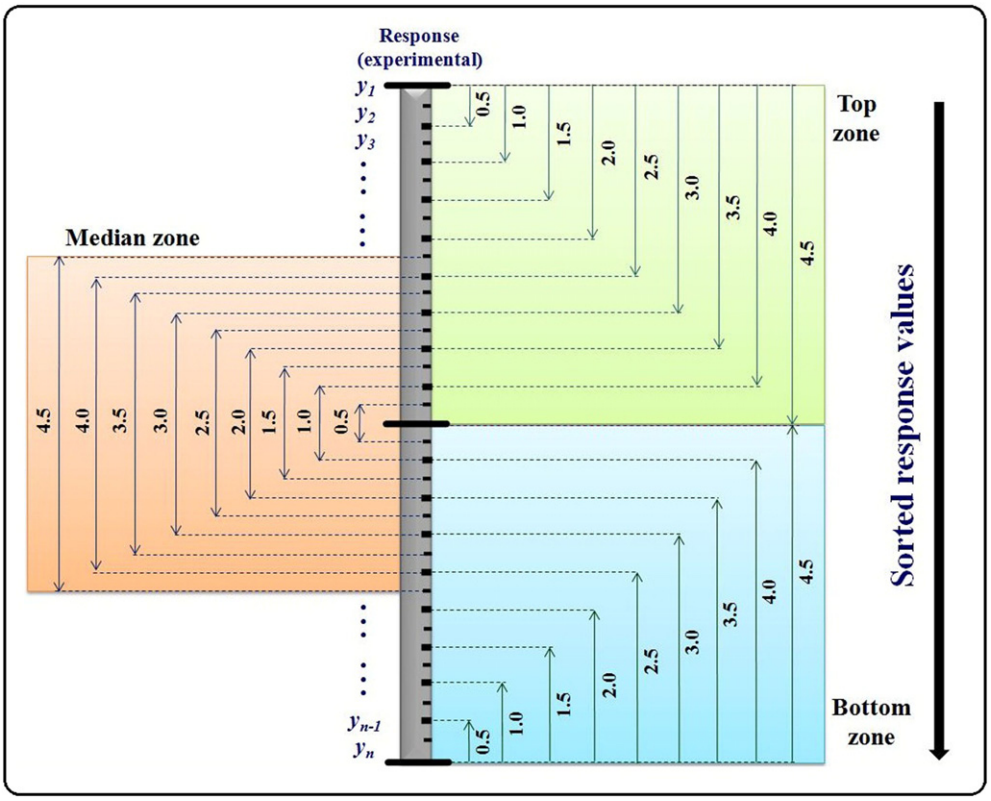

Fig. 4. The methodology implemented for the derivation of test subsets from the top, median, and bottom zones of the whole test set.

the published models for all the three datasets. The Y-randomization operation has been performed fifty times on the training set response data followed by development of linear regression based models using the same reported descriptors. Then, the total test set was used for the prediction of response values using these fifty Y-randomized models, i.e., fifty times predictions of the same test set compounds using fifty different Y-randomized models. Finally, validation metrics were computed for all the fifty randomized predictions along with the non-random model derived predictions.

4.1. Observations from individual datasets

4.1.1. Dataset 1: aqueous solubility

The whole test set of the aqueous solubility data consists of 283 chemicals with a response range of 9.130 logarithmic units. The predictions for the whole test set as well as for the individual subsets were observed to be devoid of systematic error. The performance of the model on the whole test set was found to be 'good' as per our MAE and (MAE + 3 × σ) measures which is in agreement with the judgment provided by classical metrics (R^2_{test}, $Q^2_{(extF1)}$ and $Q^2_{(extF2)}$). Thus, the results from both the MAE as well as conventional metrics are in agreement with each other in deciding the quality of model predictions for the whole test set. But the performance of the two types of metrics is not uniform for the subset divisions of the whole test set. The values of the validation metrics of original test set and all the test subsets after removing 5% of high residual data points are summarized in Table 1.

From Table 1, it is observed that the test set range and the distribution of responses remarkably affect the values of the classical validation metrics. For this dataset, it is observed that when the test set range is similar to the training set range (as in whole test set), the classical as well as MAE-based metrics show similar performance. This is due to the fact that in such case, the test set compounds cover the entire response domain similar to that covered by the training set compounds. In lower test set ranges (subsets 1–6, 8–15 and 17–21), the $Q^2_{ext(F2)}$ metric shows poor predictions (in spite of acceptable level of error values) due to the low response ranges of the test subsets. At low ranges of subsets, most of the response values are close to the test set mean and hence the $Q^2_{ext(F2)}$ metric fails to identify the good predictions. Further, the distribution of the response values also shows that most of the observations are within ± 1.0 log unit range of test subset mean. Moreover, when the training set mean is similar to the test set mean for these low range subsets, even $Q^2_{ext(F1)}$ metric fails to identify the good predictions. However, at low range test sets, MAE-based metrics correctly identify the good, moderate or bad predictions. The subsets 6, 7, 8–16, 21 and 22 are estimated as 'good' predictions by the MAE-based metrics (MAE ≤ 0.1 × training set range AND $MAE + 3 \times \sigma \leq 0.2 \times$ training set range) based on the prediction errors. By observing the distribution of errors for these subsets, the percentage of compounds having prediction errors greater than 0.25 × training set range and 0.20 × training set range are significantly low (<2.5% for most of the subsets), which are in agreement with the judgment made by the MAE-based metrics.

Further, for the subsets 1–5 and 17–20, the MAE-based metrics ($MAE < 0.1 \times$ training set range, but $MAE + 3 \times \sigma$ value is between 0.20 × training set range and 0.25 × training set range) estimated the predictions as 'moderate'. The reason for classification of these sets as moderate is the presence of a slightly higher percentage of compounds with prediction errors having magnitude larger than (0.25, 0.2, and 0.15) × training set range. The subsets (1–5) consist of 2.3–7.14% compounds having prediction errors above 0.25 × training set range (highly erroneous), 4.7–7.14% compounds with prediction errors above 0.20 × training set range (erroneous) and 9.5–14.2% compounds with prediction errors above 0.15 × training set (moderately erroneous). The subsets (17–20), comprise 0% compounds having prediction errors greater than 0.25 × training set range and 0.20 × training set range, while they contain 11–26% compounds with prediction errors greater than 0.15 × training set range. However, for all these subsets, the $Q^2_{ext(F1)}$ metric shows over-prediction. This is because the distribution of test set response values is quite apart from the training set mean.

Table 1
Results obtained from analysis of the test subsets derived using the aqueous solubility data (dataset 1). Here, training set mean = −3.403. The limiting values used are: 0.1 * training range = 0.887, 0.15 × training range = 1.331, 0.2 × training range = 1.774, 0.25 × training range = 2.218.

Sl. no.	n_{test}	Test range	Test mean	R^2 (100% data)	$Q^2_{ext(F1)}$ (100% data)	$Q^2_{ext(F2)}$ (100% data)	CCC (100% data)	$Q^2_{ext(F1)}$ (95% data)	$Q^2_{ext(F2)}$ (95% data)	MAE (100% data)	MAE + 3 * σ (100% data)	MAE (95% data)	MAE + 3 * σ (95% data)	$\%\overline{Y}_{test}$ ± 0.5	$\%\overline{Y}_{test}$ ± 1.0	$\%\overline{Y}_{test}$ ± 1.5	$\%\overline{Y}_{train}$ ± 0.5	$\%\overline{Y}_{train}$ ± 1.0	$\%\overline{Y}_{train}$ ± 1.5	$\%n_{test}$ > 0.1 × train range	$\%n_{test}$ > 0.15 × train range	$\%n_{test}$ > 0.2 × train range	$\%n_{test}$ > 0.25 × train range
Total set	283	9.130	−3.403	0.809	0.805	0.805	0.898	0.858	0.858	0.574	1.950	0.504	1.563	24.4	47.7	66.1	24.4	47.7	66.1	19.1	7.8	1.8	1.1
Top zone																							
1	14	1.460	0.033	0.324	0.903	−3.857	0.379	0.938	−2.425	0.859	2.888	0.722	2.100	57.1	100.0	100.0	0.0	0.0	0.0	42.9	14.3	7.1	7.1
2	24	1.990	−0.347	0.447	0.907	−1.573	0.515	0.941	−0.740	0.749	2.528	0.627	1.859	50.0	91.7	100.0	0.0	0.0	0.0	37.5	8.3	4.2	4.2
3	34	2.460	−0.625	0.539	0.901	−0.854	0.590	0.934	−0.338	0.690	2.447	0.590	1.863	52.9	88.2	97.1	0.0	0.0	0.0	29.4	8.8	5.9	2.9
4	57	3.000	−1.094	0.490	0.856	−0.433	0.607	0.900	−0.092	0.726	2.454	0.634	1.933	38.6	87.7	93.0	0.0	0.0	12.3	28.1	10.5	7.0	3.5
5	84	3.500	−1.487	0.552	0.829	−0.024	0.674	0.883	0.285	0.678	2.318	0.585	1.802	36.9	78.6	94.1	0.0	10.7	40.5	28.6	9.5	4.8	2.4
6	120	4.000	−1.879	0.605	0.793	0.249	0.729	0.857	0.479	0.632	2.174	0.552	1.681	35.8	75.0	91.7	7.5	37.5	58.3	23.3	8.3	4.2	2.5
7	149	4.500	−2.157	0.654	0.779	0.447	0.778	0.848	0.629	0.580	2.041	0.500	1.543	34.9	67.8	89.9	25.5	49.7	66.4	19.5	6.7	3.4	2.0
Median zone																							
8	32	0.490	−3.468	0.005	−15.042	−18.301	0.029	−9.450	−11.035	0.483	1.703	0.407	1.264	100.0	100.0	100.0	100.0	100.0	100.0	12.5	6.3	0.0	0.0
9	71	1.000	−3.488	0.272	−2.201	−2.455	0.396	−1.036	−1.154	0.429	1.519	0.366	1.149	94.4	100.0	100.0	95.8	100.0	100.0	9.9	5.6	0.0	0.0
10	103	1.480	−3.476	0.294	−0.640	−0.686	0.493	−0.093	−0.117	0.435	1.508	0.376	1.182	67.0	100.0	100.0	67.0	100.0	100.0	10.7	4.9	0.0	0.0
11	135	1.960	−3.403	0.377	−0.181	−0.181	0.580	0.179	0.179	0.470	1.676	0.409	1.329	51.1	100.0	100.0	51.1	100.0	100.0	13.3	5.9	0.7	0.7
12	161	2.480	−3.435	0.497	0.224	0.222	0.687	0.473	0.471	0.471	1.639	0.409	1.291	43.5	84.5	100.0	42.9	83.9	100.0	13.0	5.0	0.6	0.6
13	186	2.930	−3.427	0.514	0.326	0.325	0.709	0.524	0.523	0.516	1.800	0.450	1.426	37.1	72.6	100.0	37.1	72.6	100.0	16.1	6.5	1.1	1.1
14	202	3.400	−3.393	0.566	0.444	0.444	0.748	0.621	0.621	0.520	1.818	0.450	1.410	34.2	65.8	92.6	34.2	66.8	92.6	15.3	6.4	1.5	1.0
15	216	3.910	−3.380	0.606	0.533	0.533	0.776	0.669	0.669	0.527	1.818	0.462	1.442	31.5	62.0	86.6	31.9	62.5	86.6	15.3	6.5	1.4	0.9
16	230	4.490	−3.383	0.671	0.628	0.628	0.817	0.743	0.743	0.520	1.788	0.455	1.409	29.6	58.3	81.7	30.0	58.7	81.3	14.3	6.1	1.3	0.9
Bottom zone																							
17	15	1.990	−6.889	0.232	0.915	−2.703	0.327	0.925	−2.129	0.949	2.166	0.900	2.012	60.0	93.3	100.0	0.0	0.0	0.0	53.3	26.7	0.0	0.0
18	28	2.480	−6.431	0.392	0.911	−1.115	0.473	0.926	−0.728	0.808	2.164	0.747	1.971	50.0	92.9	96.4	0.0	0.0	0.0	39.3	17.9	0.0	0.0
19	40	2.940	−6.129	0.405	0.912	−0.396	0.542	0.924	−0.160	0.728	1.974	0.682	1.799	50.0	87.5	97.5	0.0	0.0	0.0	27.5	12.5	0.0	0.0
20	59	3.500	−5.720	0.546	0.894	0.078	0.660	0.912	0.220	0.677	1.973	0.627	1.777	44.1	79.7	93.2	0.0	0.0	22.0	23.7	11.9	0.0	0.0
21	86	3.990	−5.314	0.559	0.873	0.322	0.709	0.897	0.467	0.635	1.877	0.577	1.631	30.2	76.7	90.7	0.0	15.1	46.5	20.9	9.3	0.0	0.0
22	125	4.500	−4.873	0.605	0.848	0.524	0.766	0.878	0.630	0.561	1.801	0.503	1.540	32.0	68.0	91.2	17.6	41.6	63.2	18.4	8.0	0.0	0.0

The over-prediction of $Q^2_{ext(F1)}$ in the subsets can be explained from the distribution of the response values around the training set mean. For instance, in subsets 1–3 and 17–19, the percentage of test set chemicals with response values within ±0.5, ±1.0 and ±1.5 log unit range of the training set mean is 0%, which shows that the test set response values are quite far from the training set mean. Fig. 5 gives a graphical presentation of the overview of the results obtained from the aqueous solubility data. It may be observed that the classical external validation metrics $Q^2_{ext(F1)}$ and $Q^2_{ext(F2)}$ show negative values for the test subsets at lower response ranges, viz. 0.49 and 1 log unit due to the occurrence of 100% of the data around corresponding training/test set mean (Fig 5B) although the corresponding prediction error values are acceptable. Again, even if the prediction errors are relatively high as in cases of test subset ranges 1.46 and 1.99 (Fig 5B), the corresponding $Q^2_{ext(F1)}$ values are near to unity since no compounds are within the ±1 log unit of the training set mean, and the $Q^2_{ext(F2)}$ remains negative since 100% of the test set compounds reside within the ±1 log unit limit of the test set mean value. The MAE values, however, corroborate with the corresponding occurrence of high residual predictions.

Thus, the classical metrics are influenced by the test range, and they also depend on the distribution of response values around their training/test set mean. Hence, the currently suggested criteria provide a direct measure of the quality of predictions and categorize efficiently the good and bad model predictions.

The predictions obtained from Y-randomization operation applied on the training set response were of poor quality as per our proposed MAE based criteria, and these were in agreement with the corresponding $R^2_{(random)}$ values of the test sets. The results for Y-randomization can be found in Table S1 in the Supporting information section.

4.1.2. Dataset 2: AChE inhibitory activity

The whole test set for the AChE data comprises 105 compounds and it spans a range of 6.847 log units. Systematic error was observed to be absent in the predicted values for the total set as well as for the subsets. Based on the MAE and MAE + 3 × σ metrics (after omitting 5% data points with high prediction residuals), the predictions of the AChE model on the whole test set were found to be 'moderate' which is also in agreement with the judgment provided by the classical metrics (Table 2).

However, when a variety of test subsets with different ranges and distribution of response values are validated using the classical validation metrics, the results were quite aberrant. At low ranges of test subsets (subsets with range ≤ 2.5 compared to the training set range of 7.819), the classical metrics, especially $Q^2_{ext(F2)}$, behave in a quite unusual way. Here, 9 out of 21 test subsets have a range lower than or equal to 2.5. According to the MAE based metrics (checked after removal of 5% high residual observations), the predictions for low range subsets are appropriately categorized as 'good' (Table 2: subsets 9, 10, 11 and 13), 'moderate' (1, 3, 4 and 12) and 'bad' (2) based on the extent of prediction errors that are present, while it was observed that on several occasions the classical metrics were in fact failing to recognize the model performance based on the actual prediction errors. For instance, in the low range, the $Q^2_{ext(F2)}$ metric fails to recognize less erroneous sets of predictions (1, 3–5, and 9–13) and misleadingly shows values below the threshold value for these subsets. In these low range subsets, one can clearly observe that the test set mean remains very close to the individual observation values and hence the $Q^2_{ext(F2)}$ metric fails to differentiate between good predictions and the 'no model' situation. The distribution of response values for these low range test subsets clearly shows that most of the observations (>95% for almost all the low range subsets) are within ±1.0 log unit range of test subset mean. Further, in the same case, if the test set mean is similar to the training set mean (9–13), then $Q^2_{ext(F1)}$ metric also fails to differentiate between good predictions and the 'no model' situation. However, the MAE based metrics suggest that the predictions are 'moderate' for subsets 1, 3–5, and 12 (as *MAE* value < 0.1 × training set range, but *MAE* + 3 × σ value is between 0.2 × training set range and 0.25 × training set range) and 'good' for subsets 9–11 and 13 (as *MAE* value < 0.1 × training set range, and *MAE* + 3 × σ value < 0.2 × training set range). Now, the performance of the MAE-based metric estimation can be verified by observing the percentage of compounds with the prediction errors having magnitude greater than (0.25, 0.2, and 0.15) × training set range, which clearly exposes the distribution of predicted errors. Here, the subsets 1, 3–5, and 12 comprise 0% compounds having prediction error above 0.25 × training set range (highly erroneous predictions), <5% compounds having prediction errors above 0.2 × training set range, but have 11–15.9% and 25–38.46% compounds with prediction errors above 0.15 × training training set range and 0.1 × training set range, respectively and thus can be considered as 'moderate' predictions. Further, the subsets 9–11 and 13 comprise 0% compounds with prediction errors above 0.25 × training training set range, <5% compounds having prediction errors above 0.2 × training set range, while the percentages of compounds having prediction errors above 0.15 × training set range and 0.10 × training set range are significantly low (6.25–11.1% and 12–25%, respectively) and thus can be considered as 'good' predictions.

Further, the $Q^2_{ext(F1)}$ metric shows over-prediction even for low range test subsets comprising a high percentage of erroneous predictions, but they are correctly estimated as 'bad' predictions based on the MAE-based metrics. As one can observe in the test subset 2, the MAE-based criteria categorized the model performance as 'bad' based on the presence of prediction errors (*MAE* + 3 × σ > 0.25 × training set range). Further, if we see the distribution of prediction errors, the subset 2 possesses prediction errors with magnitude greater than 0.15 × training set range and 0.1 × training set range for 20.68% and 44.82% compounds, respectively; though the percentage of compounds with prediction errors >0.25 × training set range is 0%. Thus, the model predictions are judged to be 'bad' not due to the presence of high magnitude prediction errors, but due to the presence of high percentage of moderate magnitude errors. However, $Q^2_{ext(F1)}$ still shows an acceptable value (i.e., 0.671), which seems to be a wrong judgment. Such over-prediction is possible when the training set mean is significantly different from the test set mean or in other words the distribution of test set observations is significantly different from the training set mean as observed in the test subset 2. This can be verified by observing the percentage of test subset response values within ±0.5 and ±1.0 of training set mean, which are found to be notably low, i.e., 0% and 20%, respectively. Moreover, at lower ranges, the R^2_{test} based metrics (as seen in subsets 1, 3–5, and 9–13) also fail to give correct results, but the MAE based metrics (i.e., MAE and MAE + 3 × σ) are found to be steady and working properly at low ranges.

In the test subsets with considerably high ranges, the $Q^2_{ext(F1)}$ metric is highly prone to show over-prediction without reflecting the prediction errors. For instance, in subset 18, the MAE-based metrics (checked after removal of 5% high residual observations) correctly categorized the model performance as 'bad', based on the presence of prediction errors (*MAE* + 3 × σ > 0.25 × training set range). Further, the distribution of prediction errors also demonstrates the same, i.e., subset 18 possesses prediction errors with a magnitude greater than 0.2 × training set range, 0.15 × training set range and 0.1 × training set range for 20%, 20% and 50% compounds, respectively, though the percentage of compounds with prediction errors >0.25 × training set range is 0%. Thus, similar to subset 2, the model predictions are not evaluated as 'bad' because of the presence of high magnitude prediction errors, but due to the presence of high percentage of moderate magnitude errors. However, the $Q^2_{ext(F1)}$ parameter value (i.e., 0.904) clearly shows over-prediction for subset 18, despite the presence of these high prediction errors. The high values of $Q^2_{ext(F1)}$ can be elucidated from the distribution of responses of the test subset around the training set mean. For subset 18, the percentage of test set chemicals with response values within ±0.5, ±1.0 and ±1.5 log unit range of the training set mean is 0%, and within ±2.0 log unit is 40%, which clearly suggests that the distribution of test set observations

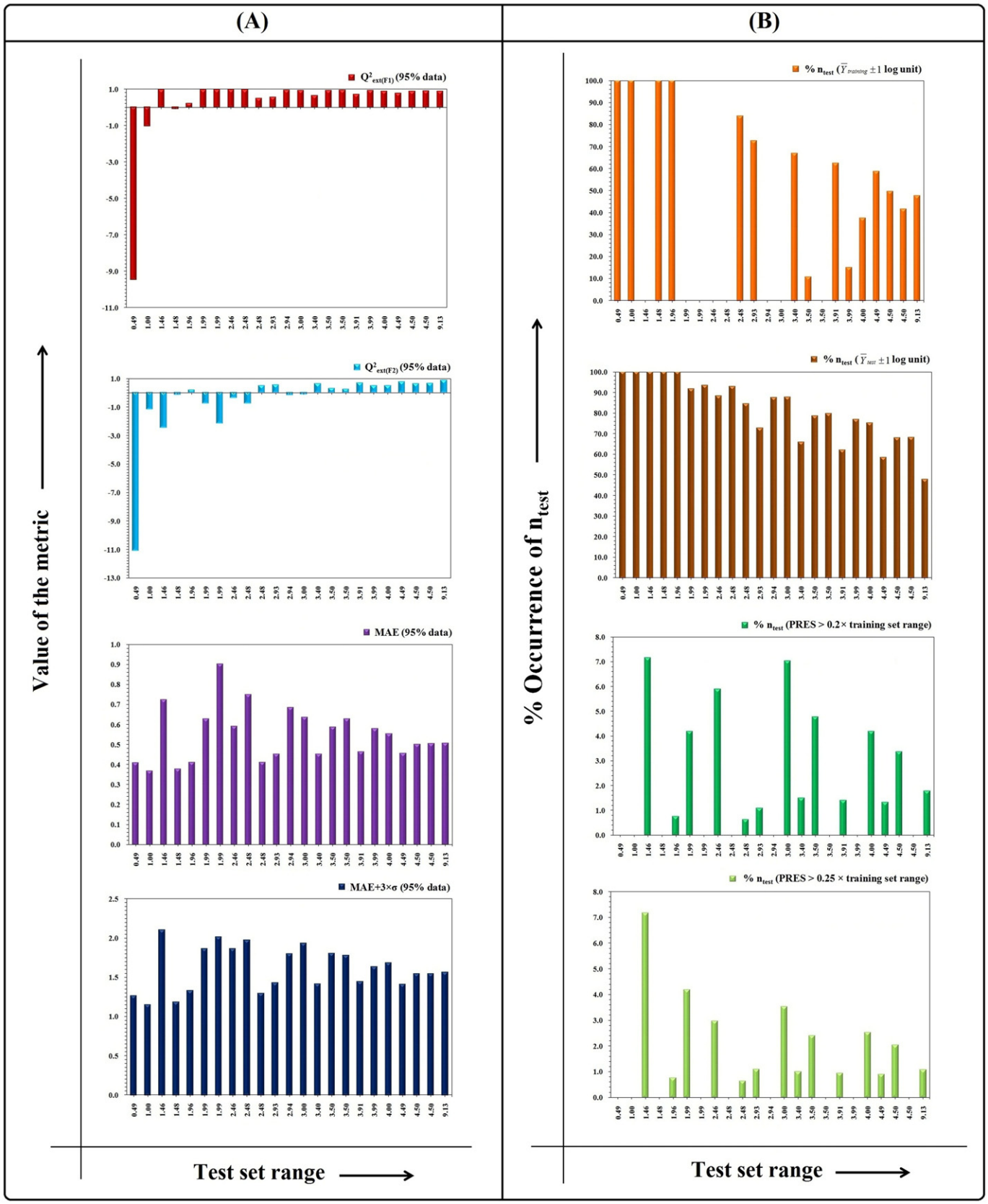

Fig. 5. Comparative analysis of selected external validation metrics using test subsets for aqueous solubility (logS) dataset: (A) Metrics computed after removal of 5% high residual data points, (B) basic data structure information for the 100% data.

Table 2
Results obtained from analysis of the test subsets derived using AChE inhibitory data (dataset 2). Here, training set mean = 6.760. The limiting values used are: 0.1 * training range = 0.780, 0.15 × training range = 1.170, 0.2 × training range = 1.560, 0.25 × training range = 1.950.

Sl. no.	n_{test}	Test range	Test mean	R^2 (100% data)	$Q^2_{ext(F1)}$ (100% data)	$Q^2_{ext(F2)}$ (100% data)	CCC (100% data)	$Q^2_{ext(F1)}$ (95% data)	$Q^2_{ext(F2)}$ (95% data)	MAE (100% data)	MAE + 3 * σ (100% data)	MAE (95% data)	MAE + 3 * σ (95% data)	$\%\overline{Y}_{test} \pm 0.5$	$\%\overline{Y}_{test} \pm 1.0$	$\%\overline{Y}_{test} \pm 1.5$	$\%\overline{Y}_{train} \pm 0.5$	$\%\overline{Y}_{train} \pm 1.0$	$\%\overline{Y}_{train} \pm 1.5$	$\%n_{test}$ > 0.1 × train range	$\%n_{test}$ > 0.15 × train range	$\%n_{test}$ > 0.2 × train range	$\%n_{test}$ > 0.25 × train range
Total set	105	6.847	6.666	0.692	0.681	0.680	0.830	0.747	0.746	0.560	1.889	0.494	1.585	26.7	52.3	82.9	25.7	52.4	84.8	28.6	10.5	3.8	0.0
Top zone																							
1	13	0.984	8.316	0.015	0.751	−7.844	0.069	0.803	−5.766	0.655	2.022	0.589	1.810	92.3	100.0	100.0	0.0	0.0	53.9	38.5	15.4	0.0	0.0
2	29	1.489	8.038	0.075	0.589	−5.706	0.171	0.671	−4.339	0.688	2.189	0.622	1.970	93.1	96.6	100.0	0.0	20.7	79.3	44.8	20.7	3.4	0.0
3	44	1.992	7.809	0.162	0.490	−2.631	0.303	0.611	−1.753	0.654	2.095	0.582	1.817	81.8	97.7	100.0	9.1	47.7	86.4	38.6	15.9	4.5	0.0
4	60	2.461	7.554	0.274	0.403	−0.778	0.467	0.539	−0.372	0.594	1.992	0.538	1.753	55.0	96.7	98.3	33.3	61.7	90.0	31.7	13.3	3.3	0.0
5	71	2.972	7.367	0.388	0.360	−0.149	0.587	0.510	0.145	0.572	1.940	0.510	1.676	47.9	88.7	98.6	38.0	67.6	91.6	29.6	12.7	2.8	0.0
6	85	3.489	7.106	0.560	0.441	0.350	0.730	0.589	0.535	0.538	1.851	0.473	1.560	38.8	68.2	96.5	31.8	64.7	92.9	27.1	10.6	2.4	0.0
7	96	3.858	6.916	0.616	0.523	0.511	0.777	0.643	0.637	0.526	1.806	0.469	1.539	30.2	59.4	94.8	28.1	57.3	92.7	26.0	9.4	2.1	0.0
8	99	4.271	6.856	0.619	0.551	0.547	0.783	0.660	0.658	0.534	1.802	0.479	1.547	28.3	55.6	90.9	27.3	55.6	89.9	27.3	9.1	2.0	0.0
Median zone																							
9	16	0.455	6.853	0.058	−7.524	−9.977	0.130	−4.237	−5.425	0.429	1.603	0.356	1.164	100.0	100.0	100.0	100.0	100.0	100.0	12.5	6.3	0.0	0.0
10	27	0.892	6.825	0.207	−5.224	−5.579	0.283	−3.420	−3.517	0.520	1.930	0.432	1.504	100.0	100.0	100.0	96.3	100.0	100.0	22.2	11.1	3.7	0.0
11	42	1.449	6.904	0.240	−0.984	−1.206	0.426	−0.261	−0.400	0.491	1.722	0.415	1.329	73.8	100.0	100.0	64.3	100.0	100.0	19.0	7.1	2.4	0.0
12	54	1.899	6.988	0.317	−0.478	−0.744	0.505	−0.096	−0.257	0.551	1.901	0.485	1.591	51.9	98.2	100.0	50.0	94.4	100.0	25.9	11.1	3.7	0.0
13	72	2.459	6.887	0.503	0.192	0.168	0.683	0.420	0.410	0.517	1.820	0.453	1.513	41.7	77.8	100.0	37.5	76.4	100.0	25.0	9.7	2.8	0.0
14	90	2.864	6.910	0.565	0.432	0.415	0.743	0.578	0.571	0.529	1.825	0.467	1.544	32.2	63.3	100.0	30.0	61.1	96.7	25.6	10.0	2.2	0.0
15	95	3.434	6.893	0.614	0.505	0.495	0.775	0.632	0.628	0.521	1.798	0.463	1.525	31.6	60.0	94.7	28.4	57.9	93.7	25.3	9.5	2.1	0.0
16	97	3.732	6.854	0.615	0.525	0.521	0.779	0.644	0.643	0.526	1.794	0.469	1.528	28.9	55.7	92.8	27.8	56.7	91.8	25.8	9.3	2.1	0.0
17	99	4.271	6.856	0.619	0.551	0.547	0.783	0.660	0.658	0.534	1.802	0.479	1.547	28.3	55.6	90.9	27.3	55.6	89.9	27.3	9.1	2.0	0.0
Bottom zone																							
18	10	2.989	4.131	0.624	0.878	−0.096	0.604	0.905	0.226	0.844	2.392	0.744	2.043	30.0	70.0	90.0	0.0	0.0	0.0	50.0	20.0	20.0	0.0
19	27	3.496	5.023	0.525	0.865	0.355	0.671	0.905	0.555	0.574	1.893	0.484	1.409	40.7	81.5	92.6	0.0	0.0	63.0	29.6	7.4	7.4	0.0
20	37	3.971	5.281	0.554	0.855	0.442	0.708	0.890	0.575	0.513	1.755	0.445	1.363	59.5	86.5	92.0	0.0	27.0	73.0	24.3	5.4	5.4	0.0
21	51	4.455	5.621	0.579	0.809	0.523	0.746	0.851	0.623	0.501	1.725	0.436	1.388	51.0	76.5	90.2	25.5	47.1	80.4	23.5	5.9	3.9	0.0

is significantly far from the training set mean and this is the only reason for the over-prediction of the metric even in the presence of high prediction errors. A graphical presentation of the results obtained from the test subsets using AChE inhibitory data is demonstrated in Fig. 6, which depicts that the $Q^2_{ext(F2)}$ values are negative if the number of compounds around test set mean ± 1 log unit is more than 90%, viz. the test subset ranges 0.46, 0.89, 0.98, and 1.49 although the MAE based criteria suggest the corresponding predictions to vary from good to moderate. Similar observations are seen for $Q^2_{ext(F1)}$ also, and interestingly, for the cases where the actual predicted error values are high, i.e., the test subset range value 2.99, the MAE based criteria provide a true judgment (bad predictions) while $Q^2_{ext(F1)}$ shows overoptimistic predictions (value 0.905).

Summarizing the observations for the AChE data, it seems clear that the MAE-based metrics (i.e., MAE and MAE + 3 × σ) work well at both low and high response ranges of the test set. Moreover, the Y-randomization analysis depicted erroneous predictions by the randomly generated models as evident from our proposed MAE based criteria and hence this imparts further confidence to the performance of the MAE based metrics. The results for Y-randomization can be found in Table S2 in the Supporting information section.

4.1.3. Dataset 3: rat cytotoxicity

Table 3 depicts the performance of various competing metrics for external validation of the dataset of ionic liquid cytotoxicity to the rat. The whole test set comprises 96 compounds covering a response range of 4.62 logarithmic units, and the predictions were observed to pass the criteria for "good predictions" based on MAE and (MAE + 3 × σ) measured after omitting 5% data points with high prediction residuals. In addition to that, other conventional external validation metrics, viz., R^2_{test}, $Q^2_{(extF1)}$ and $Q^2_{(extF2)}$ also suggest appreciable predictivity. However, consistency of various metrics varied significantly for different subsets drawn from the whole test set. From Table 3, it may be observed that two test subsets, viz. 1 and 19, show poor prediction performance while the predictions in the rest are of acceptable quality based on error judgment as per our criteria. The predictivity of the rat cytotoxicity model as determined by our proposed criteria is consistent with the presence or absence of high residual data points and is completely independent of the test set range as well as distribution of the response data. Considering 0.25 fraction of the training set range to be a limiting value for very high residuals, the two subsets 1 and 19 showing failed model performance are characterized by 14.3% and 10% of highly erroneous predictions. Three subsets 17, 20 and 21 are characterized by an MAE (at the 95% level) value being less than 0.1 × training set range, but the corresponding (MAE + 3 × σ) value falls between 0.2 × training set range and 0.25 × training set range. Hence, these three subsets do not fall under the criteria of good or bad predictions and thus they are considered to show moderate predictions. This was also in agreement with the distribution of prediction errors. In all the three cases, more than 10% of the compounds are characterized by an error magnitude of 0.15 × training set range, and subsets 20 and 21 additionally possess 6.7% and 5.3% of the highly erroneous predictions as judged by the 0.25 × training set range limit rendering such predictions to be of moderate model performance. The $Q^2_{ext(F1)}$ metric shows an overoptimistic result throughout the subsets, and even for both the cases of bad predictions (subsets 1 and 19), the value of this metric is more than 0.9 after omitting 5% high residual data points. As we can see, this is an inconsistent performance of the metric because of its inability in differentiating good and bad predictions. The high values of this metric can be explained from the distribution of responses of the test set subdivisions around the training set mean. The number of the test set chemicals with response value within ± 1 log unit range of the training set mean for subset 1 and subset 19 is 0% and in spite of the presence of a significant amount of high residuals, the predictions are overoptimistically shown to be predictive by $Q^2_{ext(F1)}$ both for 100% and 95% data. The opposite observation happens for test subsets 9 and 10 where all the test set response values (100%) are very close to the training set mean value and in spite of good model predictivity as identified by our criteria, the $Q^2_{ext(F1)}$ value fails for the 100% data, and surely this is an example of underestimation of model predictivity. However, upon removal of 5% high residual data within those subsets (subsets 9 and 10), $Q^2_{(extF1)}$ gives an acceptable value, which depicts a very low amount of prediction error among the 95% compounds allowing the metric to show good model predictivity. This observation shows that determination of the metric values after eliminating 5% high residual data points may be useful. The same problem occurs with $Q^2_{ext(F2)}$ metric where distance of the test set mean from the individual response can overshadow the actual model predictivity. Now, as $Q^2_{ext(F2)}$ uses the mean value of the test set itself, the range of the test set becomes crucial. At a low response range of the test subsets like subsets 8, 9, 10, 17 and 18, all the test subset samples (100%) lie within the ± 1 log unit range of the test set mean, and in spite of a low amount of predicted errors, the metric renders the model non-predictive for these sub-sets. However, the instance of bad predictivity is also portrayed in terms of $Q^2_{ext(F2),}$ and in two instances where actually high residual data points are present, i.e., subsets 1 and 19, the metric shows unacceptable values. At a relatively higher response range of the test subsets, $Q^2_{(extF2)}$ however shows results in agreement with our proposed criteria (Table 3). Another interesting observation occurs at subset 12 where the test set response range is 2.46 log units; for the 100% data, the metric shows a value below the threshold level, and this can be explained from the distribution of the data around test set mean which is 91.89% considering the test mean ± 1 log unit range. The value however improves for the 95% data. This is an example of prediction quality misjudgment by $Q^2_{ext(F2)}$ at a relatively higher response range even if the actual predicted errors are acceptable. Furthermore, another conventional classical parameter R^2_{test} also depicts inconsistent performance along the series and fails in true identification of the actual high prediction errors in cases like subset 1. The observations have been graphically presented in Fig. 7. It may be observed from Fig. 7 that in the lower response range of test subsets (e.g., 0.48, 0.86, 0.98 log units), the $Q^2_{ext(F2)}$ value suggests poor predictions mostly due to the occurrence of high % of chemicals close to the test set mean although $Q^2_{ext(F1)}$ shows acceptable values. In these subsets, the amount of predicted error is low but on a closer inspection at the limit of mean ± 0.5 log unit, we found that the response data are located more closely around the test set mean (i.e., $\overline{Y}_{test}$) than the training set mean (i.e., $\overline{Y}_{training}$). The higher values of the $Q^2_{ext(F1)}$ metric for the test subset ranges of 1.50 and 1.92 depict overprediction due to closeness of the data points to the training set mean (Fig. 7B) in spite of large predicted residual values which are correctly reflected in the MAE based metrics (Fig. 7A). Hence, the distribution of data around the mean plays an important role in determining the values of Q^2 based metrics while the prediction errors are truly reflected in the proposed MAE based criteria (Fig. 7A).

The randomization operation (50 times) depicted poor predictions of the whole test set with respect to our proposed MAE based criteria as well as the conventional R^2_{random} metric. The results for Y-randomization can be found in Table S3 in the Supporting information section.

Hence, we can see that various conventional metrics suffer from the influence of features like the response range and distribution of response data while judging the quality of a predictive model. Furthermore, it may be observed that the Q^2_{ext} metrics can provide different weights of judgment for a specific chemical when the composition of the test set is varied keeping that specific chemical constant. This is because Q^2_{ext} uses a mean based performance, and as the mean value changes, prediction quality of the same chemical might be judged differently. This happens in our test subsets picked from all the three regions (*vide supra*). We may mention here that we have also computed the values of another validation metric concordance correlation coefficient (CCC) [8] for all test sets/subsets (Tables 1–3). Like the R^2 based metrics, CCC also fails to provide a true reflection of actual prediction errors since its value is also affected by the test set range and the distribution of

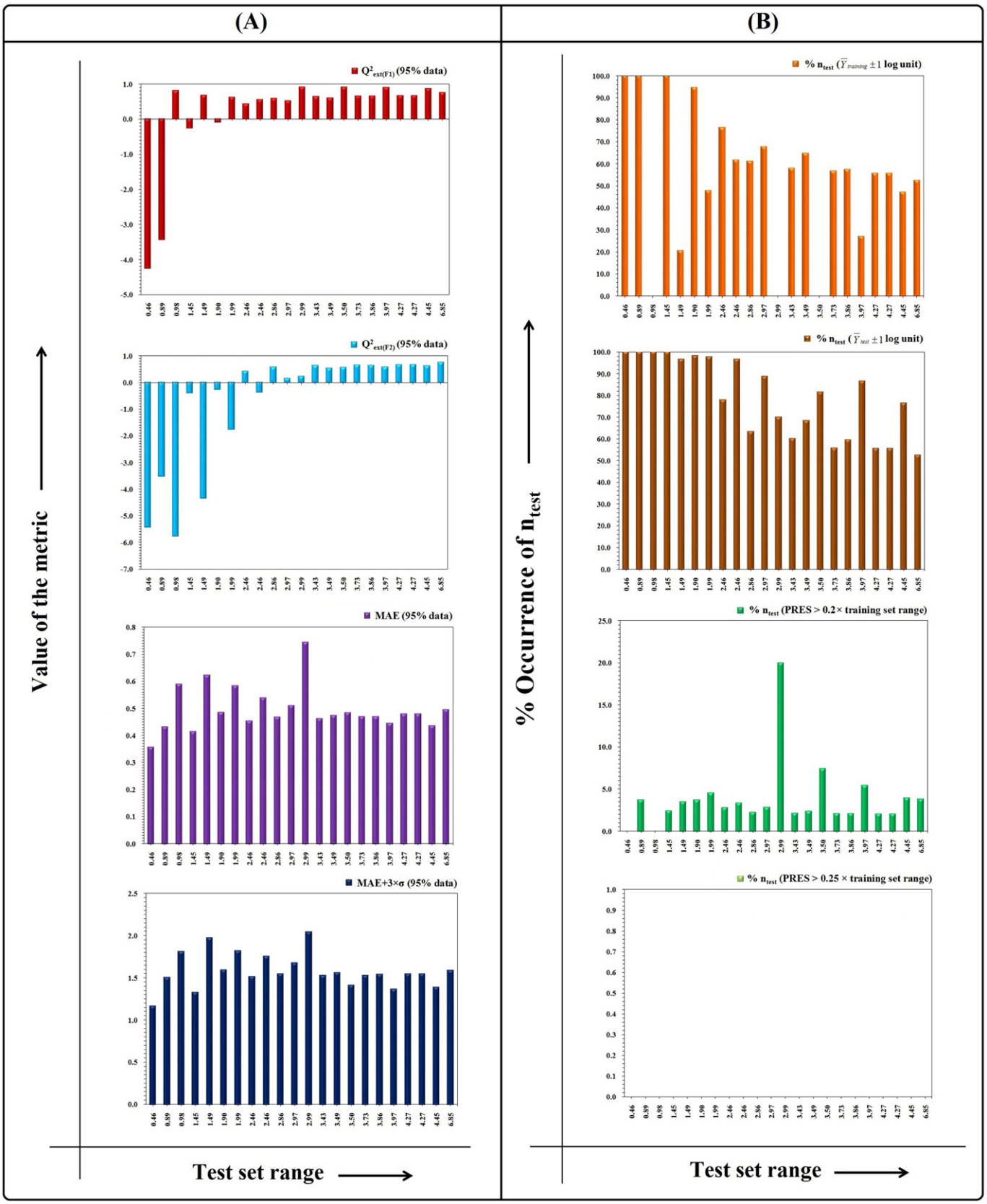

Fig. 6. Comparative analysis of selected external validation metrics using test subsets for AChE inhibitory activity dataset: (A) Metrics computed after removal of 5% high residual data points, (B) basic data structure information for the 100% data.

Table 3
Results obtained from analysis of the test subsets derived using rat cytotoxicity data (dataset 3). Here, training set mean = 3.175. The limiting values used are: 0.1 * training range = 0.500, 0.15 × training range = 0.750, 0.2 × training range = 1.000, 0.25 × training range = 1.250.

Sl. no.	n_{test}	Test range	Test mean	R^2 (100% data)	$Q^2_{ext(F1)}$ (100% data)	$Q^2_{ext(F2)}$ (100% data)	CCC (100% data)	$Q^2_{ext(F1)}$ (95% data)	$Q^2_{ext(F2)}$ (95% data)	MAE (100% data)	MAE + 3 * σ (100% data)	MAE (95% data)	MAE + 3 * σ (95% data)	$\%\overline{Y}_{test} \pm$ 0.5	$\%\overline{Y}_{test} \pm$ 1.0	$\%\overline{Y}_{test} \pm$ 1.5	$\%\overline{Y}_{train} \pm$ 0.5	$\%\overline{Y}_{train} \pm$ 1.0	$\%\overline{Y}_{train} \pm$ 1.5	$\%n_{test}$ > 0.1 × train range	$\%n_{test}$ > 0.15 × train range	$\%n_{test}$ > 0.2 × train range	$\%n_{test}$ > 0.25 × train range
Total set	96	4.620	3.232	0.819	0.816	0.815	0.903	0.894	0.894	0.351	1.384	0.292	0.964	30.2	61.5	84.4	32.3	64.6	84.4	24.0	9.4	3.1	3.1
Top zone																							
1	14	1.500	5.422	0.796	0.886	−1.478	0.616	0.938	−0.615	0.562	2.236	0.451	1.609	57.1	100.0	100.0	0.0	0.0	7.1	50.0	14.3	14.3	14.3
2	66	1.960	2.570	0.382	0.694	0.140	0.594	0.798	0.433	0.305	1.167	0.253	0.830	72.7	98.5	100.0	43.9	78.8	97.0	16.7	7.6	1.5	1.5
3	74	2.480	2.699	0.553	0.681	0.455	0.729	0.773	0.628	0.311	1.155	0.265	0.861	62.2	90.5	100.0	41.9	81.1	97.3	17.6	8.1	1.4	1.4
4	80	2.900	2.816	0.674	0.694	0.609	0.810	0.780	0.717	0.320	1.172	0.278	0.910	53.8	83.8	96.3	38.8	77.5	97.5	20.0	8.8	1.3	1.3
5	60	3.470	3.826	0.811	0.838	0.775	0.895	0.908	0.877	0.357	1.394	0.300	0.952	25.0	60.0	85.0	51.7	66.7	78.3	26.7	6.7	3.3	3.3
6	85	3.990	3.407	0.822	0.821	0.813	0.907	0.911	0.909	0.328	1.356	0.264	0.869	25.9	67.1	87.1	36.5	72.9	84.7	21.2	5.9	3.5	3.5
7	94	4.370	3.267	0.824	0.821	0.820	0.907	0.903	0.903	0.339	1.354	0.280	0.911	28.7	62.8	85.1	33.0	66.0	86.2	22.3	7.5	3.2	3.2
Median zone																							
8	58	1.390	2.677	0.358	0.609	−0.124	0.573	0.784	0.382	0.276	1.095	0.229	0.739	81.0	100.0	100.0	50.0	89.7	100.0	12.1	3.5	1.7	1.7
9	25	0.430	2.754	0.056	0.022	−10.316	−0.142	0.632	−3.540	0.297	1.273	0.223	0.666	100.0	100.0	100.0	72.0	100.0	100.0	12.0	4.0	4.0	4.0
10	43	0.860	2.737	0.190	0.402	−1.561	0.374	0.732	−0.182	0.272	1.107	0.217	0.670	100.0	100.0	100.0	58.1	100.0	100.0	11.6	2.3	2.3	2.3
11	67	1.905	2.707	0.537	0.663	0.332	0.714	0.809	0.629	0.285	1.070	0.240	0.739	68.7	97.0	100.0	46.3	85.1	100.0	13.4	3.0	1.5	1.5
12	74	2.460	2.734	0.597	0.681	0.490	0.760	0.774	0.656	0.305	1.121	0.262	0.836	59.5	91.9	100.0	41.9	82.4	98.7	17.6	6.8	1.4	1.4
13	77	2.705	2.755	0.607	0.683	0.538	0.768	0.769	0.678	0.315	1.162	0.271	0.887	57.1	89.6	100.0	40.3	80.5	97.4	19.5	7.8	1.3	1.3
14	82	3.020	2.858	0.707	0.715	0.660	0.903	0.800	0.765	0.315	1.163	0.266	0.873	51.2	81.7	95.1	37.8	75.6	97.6	19.5	8.5	1.2	1.2
15	85	3.320	2.924	0.623	0.639	0.602	0.784	0.790	0.763	0.352	1.420	0.285	0.950	47.1	78.8	94.1	36.5	72.9	95.3	22.4	10.6	3.5	3.5
16	87	3.450	2.971	0.661	0.668	0.649	0.807	0.807	0.793	0.351	1.414	0.286	0.955	46.0	77.0	92.0	35.6	71.3	93.1	23.0	10.3	3.5	3.5
Bottom zone																							
17	27	0.980	2.133	0.012	0.827	−1.930	0.069	0.872	−1.421	0.341	1.226	0.290	1.010	92.6	100.0	100.0	0.0	48.2	92.6	22.2	14.8	0.0	0.0
18	55	1.480	2.437	0.110	0.725	−0.403	0.313	0.829	0.104	0.312	1.215	0.261	0.870	83.6	100.0	100.0	32.7	74.6	96.4	16.4	9.1	1.8	1.8
19	20	1.915	5.116	0.571	0.887	−0.196	0.664	0.934	0.297	0.473	1.999	0.392	1.494	40.0	95.0	100.0	0.0	0.0	35.0	40.0	15.0	10.0	10.0
20	30	2.500	4.687	0.629	0.868	0.396	0.765	0.935	0.718	0.452	1.752	0.361	1.128	36.7	70.0	100.0	6.7	33.3	56.7	40.0	13.3	6.7	6.7
21	38	3.000	4.393	0.695	0.861	0.613	0.826	0.927	0.807	0.408	1.607	0.334	1.059	34.2	65.8	92.1	26.3	47.4	65.8	34.2	10.5	5.3	5.3
22	87	3.450	2.971	0.661	0.668	0.649	0.807	0.807	0.793	0.351	1.414	0.286	0.955	46.0	77.0	92.0	35.6	71.3	93.1	23.0	10.3	3.5	3.5
23	89	3.960	3.026	0.710	0.712	0.704	0.837	0.835	0.830	0.347	1.402	0.284	0.948	42.7	75.3	89.9	34.8	69.7	91.0	22.5	10.1	3.4	3.4
24	95	4.500	3.201	0.806	0.802	0.802	0.896	0.886	0.886	0.354	1.389	0.295	0.967	31.6	63.2	85.3	32.6	65.3	85.3	24.2	9.5	3.2	3.2

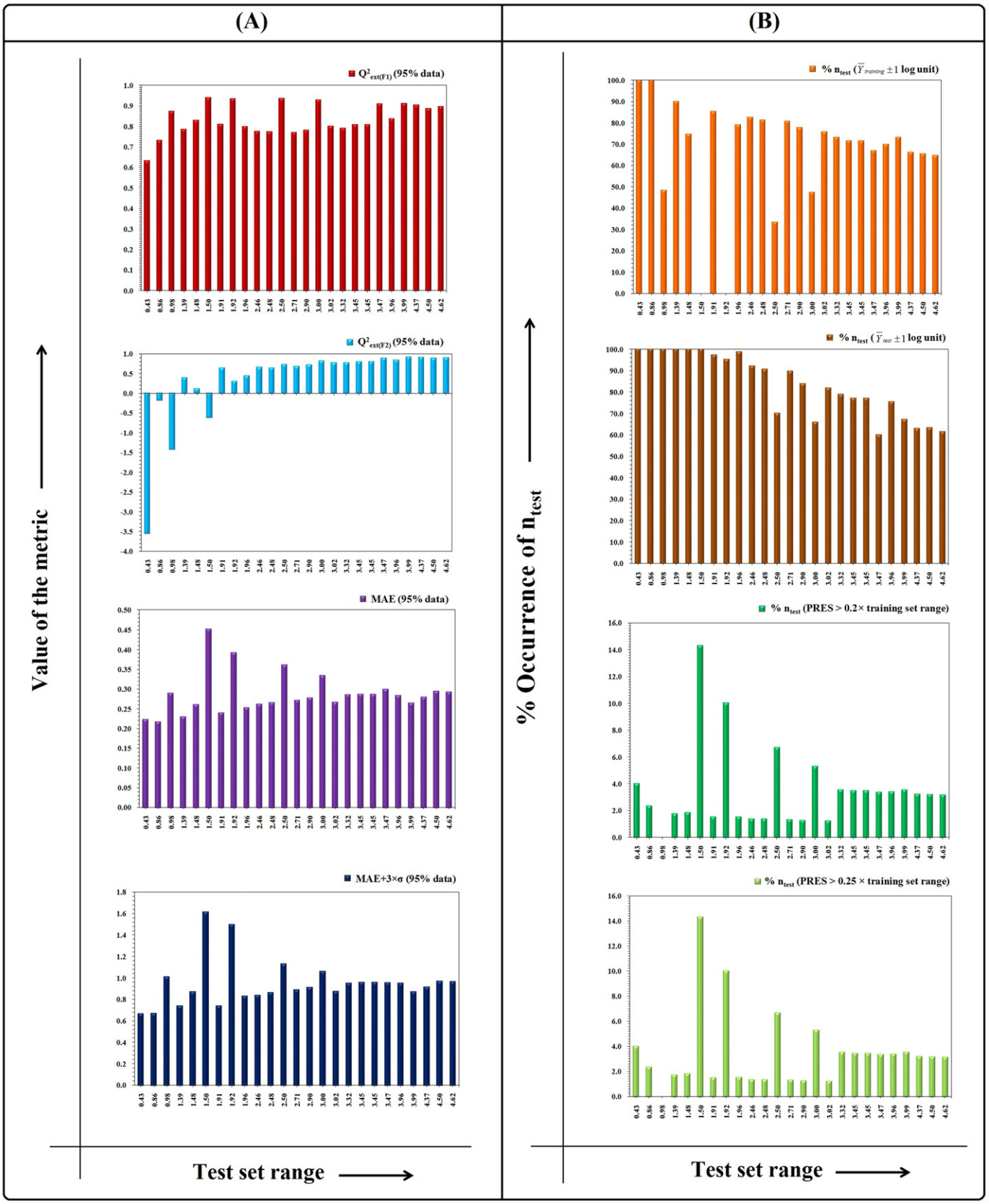

Fig. 7. Comparative analysis of selected external validation metrics using test subsets for rat cytotoxicity dataset: (A) Metrics computed after removal of 5% high residual data points, (B) basic data structure information for the 100% data.

response values around the mean. At low ranges of the test set, CCC completely fails to recognize less erroneous sets of predictions or good predictions (as observed in all the three datasets discussed in the paper) and misleadingly shows values below the recommended threshold value (<0.85) [8] for these subsets. The MAE based criteria, however, give a direct measure of prediction errors which should remain unaffected for situations like different distributions of response values around the mean. Hence, the prediction judgment provided by our proposed MAE based criteria may be considered to be more realistic than the conventional Q^2_{ext} metrics. Moreover, we believe that there should be a minimum range of response values for the test set chemicals in addition to the training set when the former is used for the validation purpose. Rather than being a representation of small subset of the whole data, the test set should cover whole response region of the training set, so that the classical R^2 based validation metrics can be used as measures of quality of predictions.

5. Conclusion

Since any statistical metric is computed from compression of a large amount of data into a single value, additional information about the dispersion of the data should be considered along with the final value of the metric. It is quite obvious that predictive QSAR models developed with any kind of experimental data (activity/property/toxicity) will have some prediction errors, and the goal of the validation process lies in judging the reliability of predictions, i.e., to identify whether the error is within allowable limits or not. In the field of QSAR, the allowable error should be such that it at least does not misclassify a compound with respect to the response. In other words, if a set of chemicals is categorized into different classes based on their response values like highly active, moderately active, and inactive, the allowable error should neither permit a compound to be within a different response class nor mis-rank the compounds among themselves based on the experimental response values. The crucial aspect of setting allowable limits for prediction errors should be very much dependent on the response range of the chemicals used in the QSAR study. We strongly believe that different validation metrics should therefore be addressed with respect to the issue of allowable error limit instead of features like 'mean based performance' as in cases of the Q^2 based metrics. For QSAR modeling analysis, use of the training set response range for the determination of allowable error limit seems very much reasonable. This means that for a training set with a wider response range, allowable prediction errors can be higher than that in the case of another training set with relatively narrow response range. It is obvious that the test set used for the purpose of validation of a given model should lie not only within the chemical domain but also within the response domain of the training set. The MAE based criteria proposed in this study consider the predictive error values based on the allowable error limit relative to the training set response range. In this way, we can actually ascertain the allowable predictive error range of a model irrespective of the size and composition of the test set.

As different model validation metrics consider summed error based feature, it is necessary to check the dispersion of errors throughout the set allowing the detection of influential observations. Hence, despite having an allowable summed prediction error value, a test set may show a significant number of high residual data points. The $MAE \pm 3 \times \sigma$ criterion enables assessment of this feature and it can render a prediction to be bad or moderate if that test set contains a significant number of data points with high prediction residuals. Furthermore, removal of a certain fraction of high residual data points (5% as proposed here) allows determination of the errors and their dispersion feature for the remaining major portion of the data and thus helps in the demonstration of model performance for most of the test set observations. In addition to that, with consideration of a specific training set and a specific test set composition, the judgment provided by the MAE based criteria after removal of 5% high residual data points enables one to choose the best model, from among some competing models, which performs with the maximum precision for most of the test set data.

Another issue we have discussed in this article is the bias of model errors. The statistical reliability of model based predictions can only be ascertained when the error values lie both in positive and negative sides of the null residual. If all or most of the errors obtained are either positive or negative showing systematically erroneous predictions, the model must be modified to remove such bias before applying for any prediction purposes.

The metrics used for the judgment of quality of a predictive model should portray the true picture of prediction errors and should not be dependent upon features like mean-based performance, range attributes or distribution pattern of data. The true applicability of a QSAR model, thus, not only depends upon the chemical domain it belongs to but also on the response domain in which it is applicable. The latter actually defines the extent of allowable prediction errors by that model. The validation approach proposed in this study is potentially useful on two basic grounds. The first one is its simplicity (it only deals with prediction errors and their distribution without any treatment or comparison with any reference condition), and the second is that it can judge/penalize a model quality directly using prediction errors with respect to the response domain of the training set. Although the inadequacy of information in the conventional R^2 based metrics is already known and error measures like MAE are commonly used in the QSAR literature, use of the MAE based criteria proposed in this study may be helpful when the conventional R^2 based metrics may be misleading for the novice users. It is thus useful to be aware (*not* beware!) of error measures and their dispersion to avoid any misleading conclusion about model predictivity which may arise from classical R^2 based validation metrics which are determined by attributes such as response range and distribution of data around the training/test set mean. The MAE based criteria suggested here along with other commonly used validation metrics may be applied to evaluate predictive performance of QSAR models with a greater degree of confidence.

Conflict of interest

None declared.

Research funding

No funding has been received for this research work.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.chemolab.2016.01.008.

References

[1] R. Benigni, A. Giuliani, Putting the predictive toxicology challenge into perspective: reflections on the results, Bioinformatics 19 (2003) 1194–1200.
[2] S. Kar, K. Roy, Predictive toxicology using QSAR: a perspective, J. Indian Chem. Soc. 87 (2010) 1455–1515.
[3] OECD environment health and safety publications series on testing and assessment No. 69, Guidance Document on the Validation of (Quantitative) Structure Activity Relationship [(Q)SAR] ModelsAvailable via 2007http://dx.doi.org/10.1787/9789264085442-en (last accessed on November 18, 2015).
[4] K. Roy, I. Mitra, On various metrics used for validation of predictive QSAR models with applications in virtual screening and focused library design, Comb. Chem. High Throughput Screen. 14 (2011) 450–474.
[5] A. Golbraikh, A. Tropsha, Beware of q 2! J. Mol. Graph. Model. 20 (2002) 269–276.
[6] A. Tropsha, Best practices for QSAR model development, validation, and exploitation, Mol. Inf. 29 (2010) 476–488.
[7] K. Roy, P. Chakraborty, I. Mitra, P.K. Ojha, S. Kar, R.N. Das, Some case studies on application of rm^2 metrics for judging quality of quantitative structure–activity relationship predictions: emphasis on scaling of response data, J. Comput. Chem. 34 (2013) 1071–1082.

[8] N. Chirico, P. Gramatica, 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. 51 (2011) 2320–2335.
[9] D.L.J. Alexander, A. Tropsha, D.A. Winkler, Beware of R 2: Simple, Unambiguous Assessment of the Prediction Accuracy of QSAR and QSPR Models, J. Chem. Inf. Model. 55 (2015) 1316–1322.
[10] W.H. Davis Jr., W.A. Pryor, Measures of goodness of fit in linear free energy relationships, J. Chem. Educ. 53 (1976) 285.
[11] S. Lothar, Z. Reynarowych, Applied Statistics: A Handbook of Techniques, Springer-Verlag, New York, 1982.
[12] K. Roy, I. Mitra, S. Kar, P.K. Ojha, R.N. Das, H. Kabir, Comparative studies on some metrics for external validation of QSPR models, J. Chem. Inf. Model. 52 (2012) 396–408.
[13] I. Mitra, P.P. Roy, S. Kar, P.K. Ojha, K. Roy, On further application of rm^2 as a metric for validation of QSAR models, J. Chemom. 24 (2010) 22–33.
[14] J. Mendham, R.C. Denney, J.D. Barnes, M.J.K. Thomas, Vogels Textbook of Quantitative Chemical Analysis, sixth ed. Pearson Education India, 2007.
[15] J.C. Dearden, M.T.D. Cronin, K.L.E. Kaiser, How not to develop a quantitative structure–activity or structure–property relationship (QSAR/QSPR), SAR QSAR Environ. Res. 20 (2009) 241–266.
[16] K.-T. Fang, H. Yin, Y.-Z. Liang, New approach by Kriging models to problems in QSAR, J. Chem. Inf. Comput. Sci. 44 (2004) 2106–2113.
[17] M.H. Abraham, R. Sanchez-Moreno, J.E. Cometto-Muniz, W.S. Cain, A quantitative structure–activity analysis on the relative sensitivity of the olfactory and the nasal trigeminal chemosensory systems, Chem. Senses 32 (2007) 711–719.
[18] V. Consonni, D. Ballabio, R. Todeschini, Comments on the definition of the Q^2 parameter for QSAR validation, J. Chem. Inf. Model. 49 (2009) 1669–1678.
[19] V. Consonni, D. Ballabio, R. Todeschini, Evaluation of model predictive ability by external validation techniques, J. Chemom. 24 (2010) 194–201.
[20] T. Chai, R.R. Draxler, Root mean square error (RMSE) or mean absolute error (MAE)? — arguments against avoiding RMSE in the literature, Geosci. Model Dev. 7 (2014) 1247–1250.
[21] C.J. Willmott, K. Matsuura, Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance, Clim. Res. 30 (2005) 79.
[22] J.H. Glick, Expression of random analytical error as a percentage of the range of clinical interest, Clin. Chem. 22 (1976) 475–483.
[23] G.W. Snedecor, W.G. Cochran, Statistical Methods, The IOWA StateUniversitypress, AMES, IOWA, VSA, 1967.
[24] R.N. Das, K. Roy, QSPR with extended topochemical atom (ETA) indices. 4. Modeling aqueous solubility of drug like molecules and agrochemicals following OECD guidelines, Struct. Chem. 24 (2013) 303–331.
[25] G. Brahmachari, C. Choo, P. Ambure, K. Roy, In vitro evaluation and in silico screening of synthetic acetylcholinesterase inhibitors bearing functionalized piperidine pharmacophores, Bioorg. Med. Chem. 23 (2015) 4567–4575.
[26] R.N. Das, K. Roy, P.L.A. Popelier, Exploring simple, transparent, interpretable and predictive QSAR models for classification and quantitative prediction of rat toxicity of ionic liquids using OECD recommended guidelines, Chemosphere 139 (2015) 163–173.