Chapter 2
Statistical Methods in QSAR/QSPR

Abstract QSAR/QSPR studies are aimed at developing correlation models using a response of chemicals (activity/property) and chemical information data in a statistical approach. The regression- and classification-based strategies are employed to serve the purpose of developing models for quantitative and graded response data, respectively. In addition to the conventional methods, various machine learning tools are also useful for QSAR/QSPR modeling analysis especially for studies involving high-dimensional and complex chemical information data bearing a nonlinear relationship with the response under consideration.

Keywords Applicability domain · Chemometric tools · Classification · MLR · Model development · OECD · Validation

2.1 Introduction

QSAR/QSPR models represent mathematical equations correlating the response of chemicals (activity/property) with their structural and physicochemical information in the form of numerical quantities, i.e., descriptors. Suitable statistical methods are deployed to derive a robust mathematical correlation involving small to large number of variables. Various regression- and classification-based methods are used for this purpose. Regression-based approaches are employed when the response data of chemicals are entirely numerical, i.e., quantitative, while qualitative or semi-quantitative chemical response(s) are modeled using classification techniques. It may be noted that the descriptors in both the cases of regression- and classification-based methods will be explicitly quantitative values. The regression-based methods enable the quantitative prediction of the response (activity/property), while classification methods allow categorization of the data points into several groups or classes such as highly active and less active. In addition to the conventional

K. Roy et al., *A Primer on QSAR/QSPR Modeling*,
SpringerBriefs in Molecular Science, DOI 10.1007/978-3-319-17281-1_2

methods, machine learning-based methods are also useful in developing QSAR/QSPR models. It may be noted that the machine learning tools employing artificial intelligence can also be used to solve regression- and classification-based problems. Now, apart from the model development formalisms, various statistical tools are also useful for feature selection from a large matrix of descriptor data. The feature selection tools enable the use of suitable and relevant descriptors for a particular response, thereby removing noises from the analysis. Furthermore, the descriptor data matrix can also be subjected to various pruning methods to reduce intercorrelated and redundant chemical information. The developed QSAR models are also subjected to several validation tests to check for the reliability of the developed correlation models. After its development, a QSAR model is usually verified by employing multiple statistical validation strategies giving an estimation of its predictivity and stability. According to the OECD guidelines, the development of a QSAR model should comply with unambiguous algorithm strategies and the model should pass various testes model fitness, robustness, and predictivity. The present chapter gives an account of various statistical tools used for the data pretreatment, feature selection, model development, and validation of QSAR/QSPR models.

2.2 Chemometric Tools

Chemometrics is the chemical discipline that uses statistical methods to design optimal procedures, experiments, and objects, and to provide maximum chemical information by analyzing chemical data.

2.2.1 Various Chemometric Tools Used in QSAR/QSPR

QSAR/QSPR is basically a statistical approach correlating the response property or activity data with descriptors encoding chemical information. Such correlation may be derived either in a regression-based approach (in cases where the response property is quantitative and available in a continuous scale) or a classification-based approach (in cases where the response property is graded or semi-quantitative).

The most commonly used regression-based approaches are as follows:

- Multiple linear regression (MLR)
- Partial least squares (PLS)

Some of the common classification-based approaches are as follows:

- Linear discriminant analysis (LDA)
- Cluster analysis

Machine learning tools such as artificial neural network, support vector machine are also very effective in developing predictive models, particularly handling with high-dimensional and complex chemical information data showing a nonlinear relationship with the response(s) of the chemicals. Some of the more popular and commonly used chemometric tools will be briefly discussed in this chapter. However, before any statistical model building method is applied, the QSAR/QSPR data table may be required to be pretreated followed by application of a suitable feature selection method.

2.2.2 Pretreatment of the Data Table

While preparing a QSAR table, care should be taken to ensure that the molecular structures have been correctly drawn or imported, the biological activity (or other response) data have been taken from an authentic source (and they have permissible experimental errors) and the descriptor values have been computed using a validated software. The response data for a QSAR modeling set should ideally have a normal distribution pattern. While clubbing two or more data sets, care must be taken to ensure that all experiments performed to determine the response values have used the same protocol. Care should also be taken to avoid duplicates in the data set. The correct tautomeric form of the structure of the compounds should also be considered. For computation of 3D descriptors, appropriate structure optimization should have been carried out.

When a large number of descriptors have been calculated, an appropriate method to remove less important or redundant descriptors should be applied. One can omit the descriptors with a constant value for all observations and the descriptors showing a very low variance. Only one descriptor among those showing high mutual intercorrelation should be retained. Descriptors showing a very low correlation with the response may also be omitted in order to thin the descriptor pool. In some cases, a suitable scaling of the descriptors may also be required.

2.2.3 Feature Selection

The selection of appropriate descriptors for model development from a pool of a large number of descriptors is an important step in QSAR modeling. Such selection may be done in a variety of ways, including stepwise selection (using a suitable stepping criterion, e.g., ‘F-for-inclusion’ and ‘F-for-exclusion’ based on partial F-statistic), all possible subset selection, genetic method, and factor analysis.

2.2.4 Multiple Linear Regression

Multiple linear regression or MLR [1] is a commonly used method in QSAR due to its simplicity, transparency, reproducibility, and easy interpretability. The generalized expression of an MLR equation will be like the following:

$$Y = a_0 + a_1 \times X_1 + a_2 \times X_2 + a_3 \times X_3 + \cdots + a_n \times X_n \tag{2.1}$$

In the above expression, Y is the response or dependent variable, X_1, X_2, …, X_n are descriptors (features or independent variables) present in the model with the corresponding regression coefficients a_1, a_2, …, a_n, respectively, and a_0 is the constant term of the model. The interpretation of contribution of individual descriptors X_1, X_2, …, X_n is straightforward depending on the corresponding coefficient value and its algebraic sign. Each regression coefficient should be significant at $p < 0.05$ which can be checked from a 't' test. The descriptors present in an MLR model should not be much intercorrelated. For a statistically reliable model, the number of observations and number of descriptors should bear a ration of at least 5:1. A MLR model that fits well the given data will lead to a scatter plot (observed vs. calculated) showing a minimum deviation of the points from the line of fit (Fig. 2.1). The quality of a MLR model is determined from a number of metrics as described below.

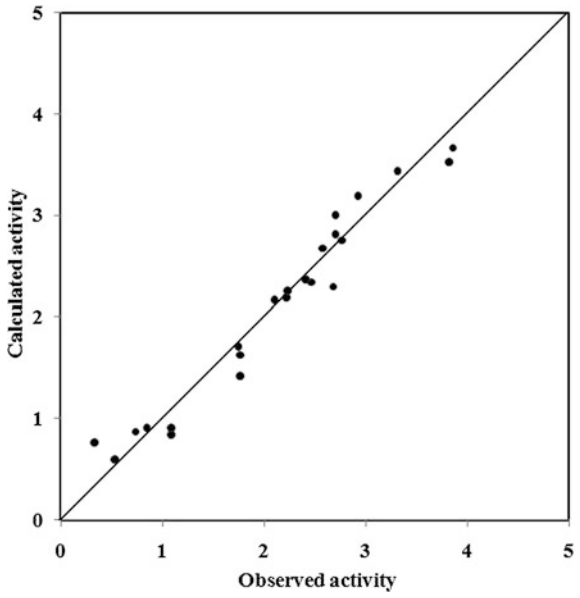

Fig. 2.1 A scatter plot of the observed and calculated activity for an MLR model

1. Determination coefficient (R^2)
   One can define the determination coefficient (R^2) in the following manner:

$$R^2 = 1 - \frac{\sum \left(Y_{\text{obs}} - Y_{\text{calc}}\right)^2}{\sum \left(Y_{\text{obs}} - \overline{Y_{\text{obs}}}\right)^2} \tag{2.2}$$

   In the above equation, Y_{obs} stands for the observed response value, while Y_{calc} is the model-derived calculated response and $\overline{Y_{\text{obs}}}$ is the average of the observed response values. For the ideal model, the sum of squared residuals being 0, the value of R^2 is 1. As the value of R^2 deviates from 1, the fitting quality of the model deteriorates. The square root or R^2 is the multiple correlation coefficient (R).
2. Adjusted R^2 (R_a^2)
   If one goes on increasing the number of descriptors in a model for a fixed number of observations, R^2 values will always increase, but this will lead to a decrease in the degree of freedom and low statistical reliability. Thus, a high value of R^2 is not necessarily as indication of a good statistical model that fits well the available data. To reflect the explained variance (the fraction of the data variance explained by the model) in a better way, adjusted R^2 which has been defined in the following manner:

$$R_a^2 = \frac{(N-1) \times R^2 - p}{N - 1 - p} \tag{2.3}$$

   In the above expression, p is the number of predictor variables used in the model development.
3. Variance ratio (F)
   To judge the overall significance of the regression coefficients, the variance ratio (the ration of regression mean square to deviations mean square) can be defined as follows:

$$F = \frac{\frac{\sum (Y_{\text{calc}} - \bar{Y})^2}{p}}{\frac{\sum (Y_{\text{obs}} - Y_{\text{calc}})^2}{N-p-1}} \tag{2.4}$$

   The F value has two degrees of freedom: p, $N - p - 1$. The computed F value of a model should be significant at $p < 0.05$. For overall significance of the regression coefficients, the F value should be high.
4. Standard error of estimate (s)
   For a good model, the standard error of estimate of Y should be low and this is defined as follows:

$$s = \sqrt{\frac{\left(Y_{\text{obs}} - Y_{\text{calc}}\right)^2}{N - p - 1}} \tag{2.5}$$

   It has a degree of freedom of $N - p - 1$.

Note that development of MLR models and computation of various statistical metrics can be done by the use of an open access tool available at http://dtclab.webs.com/software-tools and http://teqip.jdvu.ac.in/QSAR_Tools/ and also from the site http://aptsoftware.co.in/DTCMLRWeb/index.jsp.

2.2.5 Partial Least Squares (PLS)

While handling a large number of intercorrelated and noisy descriptors for a limited number of data points, PLS is a better choice over MLR. PLS, being a generalization of MLR [2], tries to extract the latent variables (LV), which are functions of the original variables, accounting for as much of the underlying factor variation as possible while modeling the responses. Before the analysis, the *X*- and *Y*-variables are often transformed to make their distributions fairly symmetrical. The response variables are usually logarithmically transformed and the *X* variables should be scaled appropriately. The linear PLS finds a few new variables (latent variables), which are linear combinations of the original variables. When the number of LVs is equal to the number of variables, the PLS model becomes same as the MLR model. A strict test of the predictive significance of each PLS component is necessary, and then stopping addition of new components when components start to be non-significant. Cross-validation (CV) is a practical and reliable way to test this predictive significance. A PLS equation can be expressed in the same form as in MLR; thus contributions of individual descriptors to the response can be easily found out.

2.2.6 Linear Discriminant Analysis

LDA [3] can separate two or more classes of objects and can thus be used for classification problems. LDA performs the same task as MLR by predicting an outcome when the response property has graded values and molecular descriptors are continuous variables. LDA explicitly attempts to model the difference between the classes of data. In a two-group situation, the predicted membership is calculated by computing a discriminant function (DF) score for each case (Fig. 2.2). Then, cases with DF values smaller than the cutoff value are classified as belonging to one group, while those with values larger are classified into the other group. The DF may take the following form:

$$\mathrm{DF} = c_1 \times X_1 + c_2 \times X_2 + \cdots + c_m \times X_m + a \tag{2.6}$$

where DF is the discriminate function, which is a linear combination (sum) of the discriminating variables, c is the discriminant coefficient or weight for that variable, X is respondent's score for that variable, a is a constant, m is the number of predictor variables. The c's are unstandardized discriminant coefficients analogous

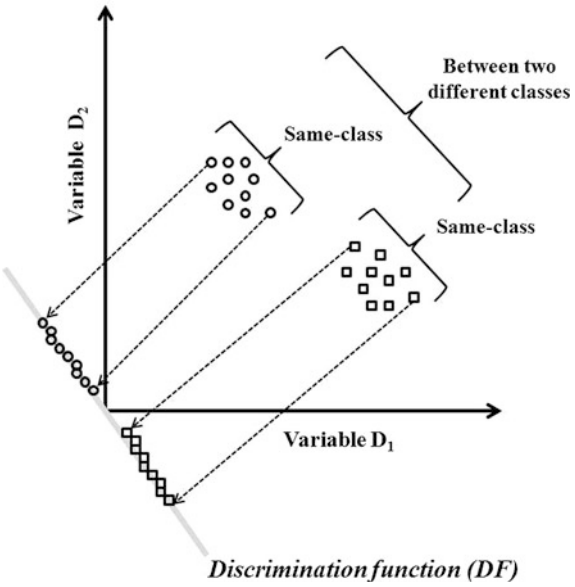

Fig. 2.2 Distribution of compounds in two groups using a discrimination function DF in a LDA analysis

to the *beta* coefficients in the regression equation. These *c*'s maximize the distance between the means of the criterion (dependent) variable. Good predictors tend to have large standardized coefficients. After using an existing set of data to calculate the DF and classify cases, any new cases (test samples) can then be classified.

In a stepwise DF analysis, the model is built step-by-step. Specifically, at each step all variables are reviewed and evaluated to determine which one will contribute most to the discrimination between groups. That variable will then be included in the model, and the process starts again.

2.2.7 Cluster Analysis

Unlike LDA, cluster analysis [4] requires no prior knowledge about which elements belong to which clusters. The clusters are defined through an analysis of the data. Cluster analysis maximizes the similarity of cases within each cluster while maximizing the dissimilarity between groups that are initially unknown.

The hierarchical cluster analysis finds relatively homogeneous clusters of cases based on dissimilarities or distances among objects. The most straightforward and generally accepted way of computing distances between objects in a multidimensional space is to compute the Euclidean distances or the squared Euclidean distance. It starts with each case as a separate cluster and then combines the clusters sequentially, reducing the number of clusters at each step until only one cluster is left. A hierarchical tree diagram or dendrogram (Fig. 2.3) can be generated to show the linkage points: the clusters are linked at increasing levels of dissimilarity.

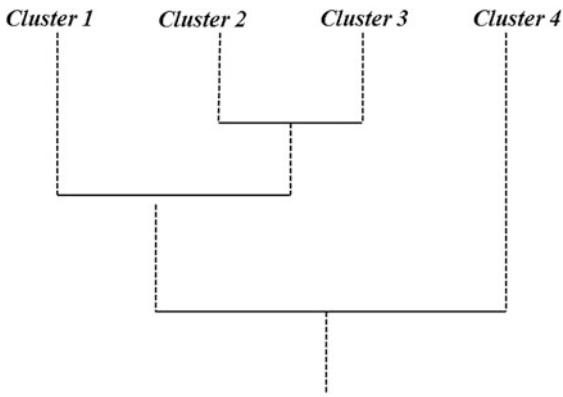

Fig. 2.3 Example of a dendrogram

The *k*-means clustering is a non-hierarchical method of clustering which can be used when the number of clusters present in the objects or cases is known. It is an unsupervised method of centroid-based clustering. In general, the *k*-means method will produce the exact *k* different clusters. The method defines *k* centroids, one for each cluster, placed as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the positions of the *k* centroids are recalculated. This procedure is repeated until the centroids no longer move.

2.3 Quality Metrics

2.3.1 Importance of Metrics for Determination of Quality of QSAR Models

Advancement in fast and economical computational resources makes it feasible to compute a large number of descriptors using various software tools. As a consequence, one cannot deny the risk of chance correlations with the increasing number of variables included in the QSAR model as compared to the limited number of compounds usually employed for the model development [5]. On the other hand, employing miscellaneous optimization tools, it is feasible to get models that can fit well the experimental data but there always remains a chance of overfitting. Fitting of data does not corroborate a good predictability of the model as the former is a parameter for the statistical quality of the model. This is the main reason why validation tools must be applied on the developed QSAR model to check its predictivity for new untested molecules. A flowchart for the method of development of a dependable QSAR model along with the various validation methods with the metrics commonly used are demonstrated in Fig. 2.4.

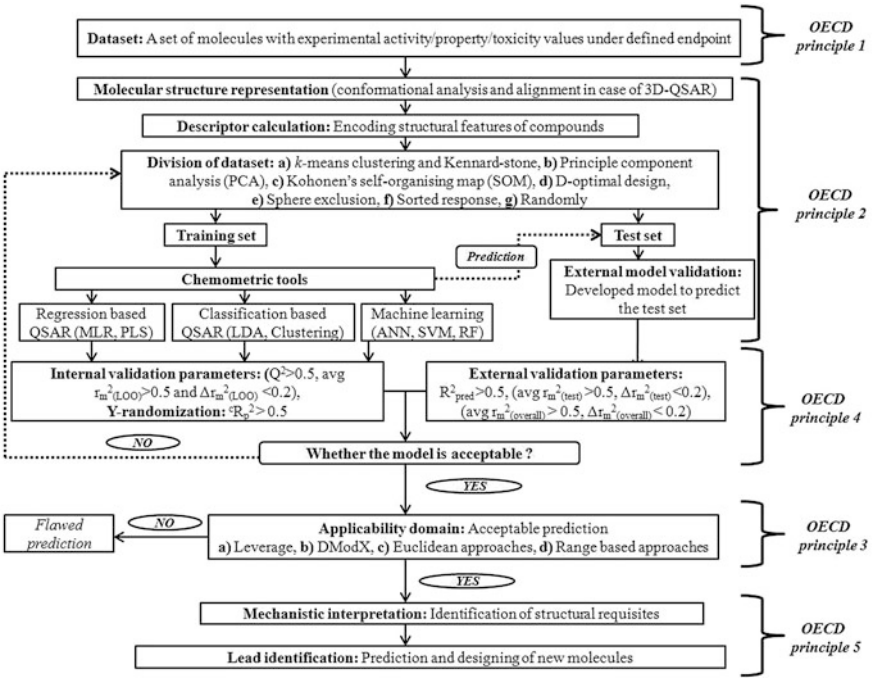

Fig. 2.4 Fundamental steps for the generation of a QSAR model and employed validation methods

2.3.2 Types of Validation

2.3.2.1 The OECD Principles

The OECD principles are the best possible outline of the essential points to be addressed while developing reliable and reproducible QSAR models [6]. The principles were formulated by QSAR experts in a meeting held in Setúbal, Portugal, in March 2002 as the guidelines for the validation of QSAR models, in particular for regulatory purposes. These principles were later approved by the OECD member countries, QSAR and regulatory communities at the 37th Joint Meeting of the Chemicals Committee and Working Party on Chemicals, Pesticides and Biotechnology in November 2004. The five guidelines adopted by the OECD denoting validity of QSAR model are as follows:

- Principle 1—A defined endpoint
- Principle 2—An unambiguous algorithm
- Principle 3—A defined domain of applicability

- Principle 4—Appropriate measures of goodness-of-fit, robustness and predictivity
- Principle 5—A mechanistic interpretation, if possible.

The present challenge in the process of development of a QSAR model is no longer in developing a model that is statistically sound to predict the activity within the training set, but in developing a model with the capability to accurately predict the activity of new chemicals.

2.3.2.2 Internal Validation

Internal validation of a QSAR model is performed based on the molecules used in the model development. It involves activity prediction of the studied molecules followed by estimation of parameters for detecting the precision of predictions. To judge the quality and goodness-of-fit of the model, internal validation is an ideal technique. But, the major disadvantage of this approach is the lack of predictability of the model when it is applied to a new data set [7].

2.3.2.3 External Validation

One cannot judge the predictability of the developed model from internal validation for an entirely new set of compounds, as internal validation considers the chemicals belonging to the same set of compounds used for model development. Thus for external validation, the available data set is usually divided into training and test sets, then subsequently a model is developed with the training set, and then the constructed model is employed to check the external validation employing the test set molecules which are not utilized in the model development process. The external validation ensures the predictability and applicability of the developed QSAR model for the prediction of untested molecules [8].

Selection of Training and Test Sets

In general, the division of the data set into training and test sets must be executed in such a manner that points representing both training and test sets are dispersed within the entire descriptor space occupied by the whole data set and each point of the test set is near to at least one compound of the training set. The following approaches are mostly employed by the QSAR practitioners for the selection of the training and test sets [8]:

1. ***Random selection***: The data set may be divided by a mere random selection process.
2. ***Based on Y-response***: This approach is based on the activity (*Y*-response) sampling. The complete range of the response is divided into bins and

compounds belonging to each bin are assigned to the training or test sets randomly or in customized way.
3. ***Based on X-response***: Properties and structural similarity of the molecules are considered for the grouping of similar compounds. After that, a predecided fraction of compounds is assigned to the training or test set manually or in some regular way.

Most commonly employed tools for the rational division of the data sets are:

- *k*-Means clustering,
- Kohonen's self-organizing map selection,
- statistical molecular design,
- Kennard–Stone selection,
- sphere exclusion, and
- extrapolation-oriented test set selection.

Note that the division of a data set using some common algorithms can be easily done by the use of an open access tool available at http://teqip.jdvu.ac.in/QSAR_Tools/.

Applicability Domain (AD)

1. Concept of the AD
   The AD is defined as a theoretical region in the chemical space constructed by both the model descriptors and modeled response. The applicability domain plays a crucial role for estimating the uncertainty in the prediction of a particular compound based on how similar it is to the compounds employed to construct the QSAR model. Therefore, the prediction of a modeled response using QSAR is applicable only if the compound being predicted falls within the AD of the model as it is unfeasible to predict the whole universe of compounds using a single QSAR model [9].
2. Types of the AD approaches
   The most commonly employed techniques for estimating interpolation regions in a multivariate space are as follows:

   (a) Ranges in the descriptor space,
   (b) geometrical methods,
   (c) distance-based methods,
   (d) probability density distribution, and
   (e) range of the response variable.

The first four approaches are based on the methodology used for interpolation space characterization in the model descriptor space. On the contrary, the last one depends solely on response space of the training set molecules. A compound can be identified out of the AD, if: (a) at least one descriptor is out of range for the ranges approach and (b) the distance between the chemical and the center of the training

data set exceeds the threshold for distance approaches. The threshold for all kinds of distance methods is the largest distance between the training set data points and the center of the training data set.

2.3.3 Validation Metrics for Regression-Based QSAR Models

2.3.3.1 Metrics for Internal Validation

The most commonly employed internal metrics are discussed below [10]:

1. Leave-one-out (LOO) cross-validation
   To determine the LOO cross-validation, the training set is primarily modified by eliminating one compound from the set. The QSAR model is then rebuilt based on the remaining molecules of the training set using the descriptor combination originally selected, and the activity of the deleted compound is computed based on the resulting QSAR equation. This cycle is repeated until all the molecules of the training set have been deleted once, and the predicted activity data obtained for all the training set compounds are used for the calculation of various internal validation parameters. Finally, the model predictivity is judged using the predicted residual sum of squares (PRESS) and cross-validated R^2 (Q^2) for the model while the value of standard deviation of error of prediction (SDEP) is calculated from PRESS.

$$\text{PRESS} = \sum \left(Y_{\text{obs}} - Y_{\text{pred}}\right)^2 \tag{2.7}$$

$$\text{SDEP} = \sqrt{\frac{\text{PRESS}}{n}} \tag{2.8}$$

$$Q^2 = 1 - \frac{\sum \left(Y_{\text{obs(train)}} - Y_{\text{pred(train)}}\right)^2}{\sum \left(Y_{\text{obs(train)}} - \overline{Y}_{\text{training}}\right)^2} = 1 - \frac{\text{PRESS}}{\sum \left(Y_{\text{obs(train)}} - \overline{Y}_{\text{training}}\right)^2} \tag{2.9}$$

   In Eqs. (2.7)–(2.9), Y_{obs} and Y_{pred} correspond to the observed and LOO-predicted activity values, n refers to the number of observations, $Y_{\text{obs(train)}}$ is the observed activity, $Y_{\text{pred(train)}}$ is the predicted activity of the training set molecules based on the LOO technique. The threshold value of Q^2 is 0.5.
2. Leave-many-out (LMO) cross-validation
   The basic principle of the LMO technique or leave-some-out (LSO) technique is that a definite portion of the training set is held out and eliminated in each cycle. For each cycle, the model is constructed based on the remaining molecules (and using the originally selected descriptors) and then the activity of the deleted

compounds is predicted using the developed model. After all the cycles have been completed, the predicted activity values of the compounds are used for the calculation of the LMO-Q^2.

3. True Q^2
   Hawkins et al. [11] proposed the concept of 'true Q^2' parameter, calculated based on application of the variable selection strategy at each validation cycle. The parameter may be a better tool for assessing model predictivity, chiefly in the case of small data sets, compared to the traditional approach of the splitting of the data set into training and test sets.
4. The ${r_m}^2$ metric for internal validation
   An acceptable value of Q^2 does not inevitably indicate that the predicted activity data lie in close propinquity to the observed ones although there may exist a good overall correlation between the values. Thus, to obviate this problem and to better indicate the model predictability, the ${r_m}^2$ metrics introduced by Roy et al. [12] may be computed by the following equations:

$$\overline{{r_m}^2} = \frac{\left({r_m}^2 + {r'_m}^2\right)}{2} \tag{2.10}$$

$$\Delta {r_m}^2 = \left|{r_m}^2 - {r'_m}^2\right| \tag{2.11}$$

Here, ${r_m}^2 = r^2 \times \left(1 - \sqrt{\left(r^2 - r_0^2\right)}\right)$ and $r{r'_m}^2 = r^2 \times \left(1 - \sqrt{\left(r^2 - {r'_0}^2\right)}\right)$. The parameters r^2 and r_0^2 are the squared correlation coefficients between the observed and (leave-one-out) predicted values of the compounds with and without intercept, respectively. The parameter ${r'_0}^2$ bears the same meaning but uses the reversed axes.

The $\overline{{r_m}^2}$ is the average value of ${r_m}^2$ and ${r'_m}^2$, and Δr_m^2 is the absolute difference between ${r_m}^2$ and ${r'_m}^2$. In case of internal validation of the training set, the $\overline{{r_m}^2{}_{(\mathrm{LOO})}}$ and $\Delta {r_m}^2{}_{(\mathrm{LOO})}$ parameters can be employed and it has been shown that the value of $\Delta {r_m}^2{}_{(\mathrm{LOO})}$ should preferably be lower than 0.2 provided that the value of $\overline{{r_m}^2{}_{(\mathrm{LOO})}}$ is more than 0.5. Roy et al. [13] proposed that the calculation of the r_m^2 metrics should be based on the scaled values of the observed and the predicted response data. The scaling may be done based on the following equation.

$$\text{Scaled } Y_i = \frac{Y_i - Y_{\min(\mathrm{obs})}}{Y_{\max(\mathrm{obs})} - Y_{\min(\mathrm{obs})}} \tag{2.12}$$

Here, Y_i refers to the observed/predicted response for the ith (1, 2, 3, …, n) compound in the training/test set. Besides these, $Y_{\max(\mathrm{obs})}$ and $Y_{\min(\mathrm{obs})}$ indicate the maximum and minimum values, respectively, for the observed response in the training set compounds.

To make the calculation of r_m^2 metrics easier, a web application known as 'r_m^2 calculator' (http://aptsoftware.co.in/rmsquare) has been also developed.

5. True $r_{m\ (\mathrm{LOO})}^2$
   In case of LOO-CV, r_m^2 is calculated based on the LOO-predicted activity values of the training set and the parameter is referred to as $r_{m\ (\mathrm{LOO})}^2$, while the true $r_{m\ (\mathrm{LOO})}^2$ value is obtained from the model developed from the undivided data set after the application of variable selection strategy at each cycle of validation [14]. The 'true $r_{m\ (\mathrm{LOO})}^2$' metric may reflect characteristics of external validation without loss of chemical information.
6. Metrics for chance correlation: *Y*-randomization
   Y-randomization is performed in order to ensure the robustness of the developed QSAR model. In the *Y*-randomization test, validation is performed by permuting the response values (Y) with respect to the X matrix which has been kept unaltered. This method is generally performed in two different ways: (a) process randomization and (b) model randomization performed at varying confidence levels. The deviation in the values of the squared mean correlation coefficient of the randomized model (R_r^2) from the squared correlation coefficient of the non-random model (R^2) is reflected in the value of ${}^cR_p^2$ parameter computed from the following equation [15]:

$$ {}^cR_p^2 = R \times \sqrt{R^2 - R_r^2} \tag{2.13} $$

The threshold value of ${}^cR_p^2$ is 0.5. For a QSAR model having the corresponding value above the stated limit, it might be considered that the model is not obtained by chance only.

2.3.3.2 Metrics for External Validation

1. Predictive R^2 $\left(R_{\mathrm{pred}}^2 \text{ or } Q_{(F1)}^2\right)$
   The R_{pred}^2 reflects the degree of correlation between the observed and predicted activity data of the test set.

$$ R_{\mathrm{pred}}^2 = 1 - \frac{\sum \left(Y_{\mathrm{obs(test)}} - Y_{\mathrm{pred(test)}}\right)^2}{\sum \left(Y_{\mathrm{obs(test)}} - \overline{Y}_{\mathrm{training}}\right)^2} \tag{2.14} $$

Here, $Y_{\mathrm{obs(test)}}$ and $Y_{\mathrm{pred(test)}}$ are the observed and predicted activity data for the test set compounds, while $\overline{Y}_{\mathrm{training}}$ indicates the mean observed activity of the training set molecules. Thus, models with values of R_{pred}^2 above the stipulated value of 0.5 are considered to be well predictive.

2. Golbraikh and Tropsha's criteria
   Golbraikh and Tropsha [16] proposed a set of parameters for determining the external predictability of QSAR model. According to Golbraikh and Tropsha, models are considered satisfactory, if all of the following conditions are satisfied:

   (a) $Q^2_{\text{training}} > 0.5$.
   (b) $R^2_{\text{test}} > 0.6$.
   (c) $\dfrac{r^2 - r_0^2}{r^2} < 0.1 \quad \text{and} \quad 0.85 \leq k \leq 1.15 \quad \text{or}$
   $\dfrac{r^2 - r'^2_0}{r^2} < 0.1 \quad \text{and} \quad 0.85 \leq k' \leq 1.15.$
   (d) $\left|r_0^2 - {r'}_0^2\right| < 0.3.$

   The meaning of the r^2 and r_0^2 terms is already discussed in the 'r^2_m *metric for internal validation*' section.
3. The $r^2_{m(\text{test})}$ metric for external validation
   In order to verify the propinquity between the observed and predicted data, the parameter $r^2_{m(\text{test})}$, similar to $r^2_{m(\text{LOO})}$ used in internal validation, has been developed by Roy et al. [12]. The value of $r^2_{m(\text{test})}$ is calculated using the squared correlation coefficients between the observed and predicted activity of the test set compounds. For the acceptable prediction, the value of $\Delta r^2_{m\ (\text{test})}$ should preferably be lower than 0.2 provided that the value of $\overline{r^2_{m(\text{test})}}$ is more than 0.5. More interestingly, Roy and coworkers established that this tool can be extended to the entire data set employing the LOO-predicted activity for the test set compounds. These parameters have been referred to as $\overline{r^2_{m(\text{overall})}}$ and $\Delta r^2_{m\ (\text{overall})}$ which reflect the predictive ability of the model for the entire data set.
4. RMSEP
   External predictive ability of a QSAR model may further be determined by root mean square error in prediction (rmsep) given by Eq. (2.15).

$$\text{RMSEP} = \sqrt{\frac{\sum \left(y_{\text{obs(test)}} - y_{\text{pred(test)}}\right)^2}{n_{\text{ext}}}} \tag{2.15}$$

   Here, n_{ext} refers to the number of test set compounds.
5. $Q^2_{(F2)}$
   $Q^2_{(F2)}$ is based on prediction of test set compounds ($Q^2_{(F2)}$) proposed by Schüürmann et al. [17] as given by Eq. (2.16).

$$Q^2_{(F2)} = 1 - \frac{\sum \left(Y_{\text{obs(test)}} - Y_{\text{pred(test)}}\right)^2}{\sum \left(Y_{\text{obs(test)}} - \overline{Y}_{\text{test}}\right)^2} \tag{2.16}$$

Here, $\bar{Y}_{\text{test}}$ refers to the mean observed data of the test set compounds. A threshold value 0.5 is defined for this parameter.

6. $Q^2_{(F3)}$

The $Q^2_{(F3)}$ metric with a threshold value of 0.5, for validation of a QSAR model has been proposed by Consonni et al. [18]. This parameter is defined as follows:

$$Q^2_{(F3)} = 1 - \frac{\left[\sum \left(Y_{\text{obs(test)}} - Y_{\text{pred(test)}}\right)^2\right] \Big/ n_{\text{ext}}}{\left[\sum \left(Y_{\text{obs(train)}} - \bar{Y}_{\text{train}}\right)^2\right] \Big/ n_{\text{tr}}} \tag{2.17}$$

where n_{tr} refers to the number of compounds in the training set. However, although the value of $Q^2_{(F3)}$ measures the model predictability, it is sensitive to training set data selection and tends to penalize models fitted to a very homogeneous data set even if predictions are close to the truth.

7. Concordance correlation coefficient (CCC)

The CCC parameter can be calculated in order to check the model reliability by the following equation [19]:

$$\overline{\rho}_c = \frac{2\sum_{i=1}^{n}\left(x_{\text{obs(test)}} - \overline{x_{\text{obs(test)}}}\right)\left(y_{\text{pred(test)}} - \overline{y}_{\text{pred(test)}}\right)}{\sum_{i=1}^{n}\left(x_{\text{obs(test)}} - \overline{x_{\text{obs(test)}}}\right)^2 + \sum_{i=1}^{n}\left(y_{\text{pred(test)}} - \overline{y_{\text{pred(test)}}}\right)^2 + n\left(\overline{x_{\text{obs(test)}}} - \overline{y_{\text{pred(test)}}}\right)} \tag{2.18}$$

In the above equation, $x_{\text{obs(test)}}$ and $y_{\text{pred(test)}}$ correspond to the observed and predicted values of the test compounds, n is the number of chemicals, and $\overline{x_{\text{obs(test)}}}$ and $\overline{y_{\text{pred(test)}}}$ correspond to the averages of the observed and predicted values, respectively, for the test compounds. The ideal value of CCC should be equal to 1.

The $r^2_{m(\text{rank})}$ metric

In order to assess the closeness between the order of the predicted activity and that of the observed activity, the $r^2_{m(\text{rank})}$ parameter was developed. The $r^2_{m(\text{rank})}$ metric is computed based on the correlation of the ranks generated for the observed and the predicted response data. An ideal ranking where the observed and the predicted response data perfectly match with each other yields zero difference between the two values for each molecule, and the $r^2_{m(\text{rank})}$ metric attains a value of unity.

$$r^2_{m(\text{rank})} = r^2_{(\text{rank})} \times \left(1 - \sqrt{r^2_{(\text{rank})} - r^2_{0(\text{rank})}}\right) \tag{2.19}$$

2.3.4 Validation Metrics Employed in Classification-Based QSAR

Validation metrics can assess the performance of the classification-based models in terms of accurate qualitative prediction of the dependent variable. Commonly applied metrics for classification-based QSAR models are illustrated below [20]:

2.3.4.1 Parameters for Goodness-of-Fit and Quality Determination

1. Wilks lambda (λ) statistics
   The Wilks lambda is a metric for the testing of significance of a discriminant model function and determined as the ratio of within group sum of squares and total sum of squares, i.e., within-category to total dispersion.

$$\text{Wilks } \lambda = \frac{\text{Within group sum of squares}}{\text{Total sum of squares}} \tag{2.20}$$

   The Wilks lambda value spans from 0 to 1, where 0 corresponds to good level of discrimination and 1 refers to no discrimination.
2. Canonical index (R_c)
   The quantification of the strength of the relationship between the dependent and independent variables is articulated as a canonical correlation coefficient.

$$R_c = \sqrt{\frac{\lambda_i}{1+\lambda_i}} \tag{2.21}$$

   Here, λ_i is referred as eigen value of the matrix.
3. Chi-square (χ^2)
   The quality of classification-based model is also judged using the chi-square (χ^2) statistic.

$$\chi^2 = \sum_{i=1}^{t} \frac{(f_i - F_i)^2}{F_i} \tag{2.22}$$

   where f_i is observed response, F_i is predicted response, and t is the number of observations.
4. Squared Mahalanobis distance
   The square of Mahalanobis distance is calculated for the determination of probability of a compound to be classified in a definite group in the discriminant

space for LDA. In a multivariate normal distribution with covariance matrix Σ, the Mahalanobis distance between any two data points x_i and x_j can be defined as follows:

$$d_{\text{mahalanobis}}(x_i, x_j) = \sqrt{(x_i - x_j)^T \sum^{-1} (x_i - x_j)} \tag{2.23}$$

where x_i and x_j are two random data points, T is transpose of a matrix, and Σ^{-1} is inverse of the covariance matrix.

2.3.4.2 Metrics for Model Performance Parameters

1. Sensitivity, Specificity, and Accuracy
   The compounds classified employing the classification-based QSAR model can be divided into four categories based on a comparison between the predicted and observed response:
   (a) True positives (TP): the active compounds which have been correctly predicted as actives,
   (b) False negatives (FN): this class includes the active compounds which have been erroneously classified as inactives,
   (c) False positives (FP): this class comprises the inactive compounds wrongly classified as actives,
   (d) True negatives (TN): this class accounts for the inactive compounds which have been accurately predicted as inactives.
   Based on the two-by-two confusion matrix, the following metrics can be computed to evaluate the classifier model performance and classification capability.

$$\text{Sensitivity} = \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2.24}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{2.25}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{2.26}$$

2. *F*-measure and Precision
   The *F*-measure refers to the harmonic mean of recall and precision, where recall refers to the accuracy of real prediction and precision defines the accuracy of a predicted class.

$$F\text{-measure} = \frac{2(\text{Recall})(\text{Precision})}{\text{Recall} + \text{Precision}} \tag{2.27}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} = \text{fp rate} \tag{2.28}$$

3. *G*-means
   Combining sensitivity and specificity into a single parameter via the geometric mean (*G*-means) allows for a straightforward way to assess the model's ability to perfectly classify active and inactive samples using the formula:

$$G\text{-means} = \sqrt{\text{Sensitivity} \times \text{Specificity}} \tag{2.29}$$

4. Cohen's κ
   Cohen's kappa (κ) can be employed to determine the agreement between classification (predicted) models and known classifications. It can be defined as follows:

$$\text{Cohen's } \kappa = \frac{P_r(a) - P_r(e)}{1 - P_r(e)} \tag{2.30}$$

$$P_r(a) = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{FP} + \text{FN} + \text{TN})} \tag{2.31}$$

$$P_r(e) = \frac{\{(\text{TP} + \text{FP}) \times (\text{TP} + \text{FN})\} + \{(\text{TN} + \text{FP}) \times (\text{TN} + \text{FN})\}}{(\text{TP} + \text{FN} + \text{FP} + \text{TN})^2} \tag{2.32}$$

   Here, $P_r(a)$ is the relative observed agreement between the predicted classification of the model and the known classification, and $P_r(e)$ is the hypothetical probability of chance agreement. Cohen's kappa analysis returns values between −1 (no agreement) and 1 (complete agreement).
5. Matthews correlation coefficient (MCC)
   The MCC is regarded as a balanced measure which can be employed even if the classes are of diverse sizes. The MCC is simply a correlation coefficient between the observed and predicted binary classifications, and it returns a value between −1 and +1. A coefficient of +1 signifies a perfect prediction, 0 an average random prediction, and −1 an inverse prediction. The MCC can be computed directly from the confusion matrix using the formula:

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}} \tag{2.33}$$

The meaning of TP, TN, FP, and FN are already discussed.

2.3.5 *Parameters for Receiver Operating Characteristics (ROC) Analysis*

1. ROC curve
   The ROC curve is a visual illustration of the success and error observed in a classification model. The curve is plotted taking true positive rate (tp) on the *y*-axis and false-positive rate (fp) on the x-axis, and the characteristics of the curve provides easier recognition of the precision of prediction [21].

$$\text{tp rate} \approx \frac{\text{Positives (active molecules) correctly classified}}{\text{Total postives}} = \text{Sensitivity} \tag{2.34}$$

$$\begin{aligned}\text{fp rate} &= \frac{\text{Negatives (inactive compounds) incorrectly classified}}{\text{Total negatives}} \\ &= 1 - \text{specificity}\end{aligned} \tag{2.35}$$

   The ROC curve signifies the number of objects the classifier identifies correctly as well as the number wrongly identified by the classifier.
2. ROCED and ROCFIT
   Two metrics based on distances in a ROC curve for the selection of classification models with an correct balance in both training and test sets, namely the ROC graph Euclidean distance (ROCED) and the ROC graph Euclidean distance corrected with fitness function (FIT(λ)) or Wilks λ (ROCFIT), are also used [22]. The Euclidean distance between the perfect and a real classifier (d_i) expressed as a function of their respective values of sensitivity and specificity is

$$d_1 = \sqrt{\left(\mathrm{Se_p} - \mathrm{Se_r}\right)^2 + \left(\mathrm{Sp_p} - \mathrm{Sp_r}\right)^2} \tag{2.36}$$

   where $\mathrm{Se_p}$ and $\mathrm{Se_r}$ are the respective sensitivity values of the perfect and the real classifier, while $\mathrm{Sp_p}$ and $\mathrm{Sp_r}$ represent the specificity values of the perfect and real classifier, respectively. Since the sensitivity and specificity for a perfect classifier takes values of 1, the Euclidean distance can be expressed as

$$d_1 = \sqrt{(1 - \mathrm{Se_r})^2 + (1 - \mathrm{Sp_r})^2} \tag{2.37}$$

$$\mathrm{ROCED} = (|d_1 - d_2| + 1) \times (d_1 + d_2) \times (d_2 + 1) \tag{2.38}$$

where d_1 and d_2 are representation of the distances in a ROC graph for the training and test sets, respectively. ROCED takes values between 0 (perfect classifier) and 4.5 (random classifier).
A new parameter ROCFIT has also been introduced. ROCFIT is defined as follows:

$$\mathrm{ROCFIT} = \frac{\mathrm{ROCED}}{\mathrm{Wilks}(\lambda)} \tag{2.39}$$

2.3.5.1 Metrics for Pharmacological Distribution Diagram (PDD)

The PDD is a frequency distribution plot of a dependent variable where expectancy values of the variable are plotted in the *y*-axis against numeric intervals of the variable in the *x*-axis [23]. This graph visually signifies the overlapping regions of the categories, e.g., positives and negatives. For a classification case comprising two classes such as actives and inactives (or positives and negatives), two terms named 'active expectancy' and 'inactive expectancy' may be defined as below where the denominator is added with a numerical value of 100 to avoid division by zero:

$$\text{Activity expectancy} = E_a = \frac{\text{Percentage of actives}}{(\text{Percentage of inactives} + 100)} \tag{2.40}$$

$$\text{Inactivity expectancy} = E_i = \frac{\text{Percentage of inactives}}{(\text{Percentage of actives} + 100)} \tag{2.41}$$

where '*a*' and '*i*' are the number of occurrences of active and inactive compounds at a specific range.

2.4 Conclusion

The QSAR/QSPR modeling technique involves the use of a significant number of statistical tools and hence requires a good knowledge of chemometrics. The developed QSAR model can furnish linear as well as nonlinear relationship between the response and chemical attributes through regression-based as well as classification-based analyses. Since, quantitative mathematical relationships are

established, validation of the models using a suitable statistical algorithm becomes essential to confirm the stability and predictivity of the models. The judgment for the choice of method depends upon a multitude of factors including the response to be modeled, the nature of the training set data, the type of descriptors used and also its numbers, and even the objective of the analysis.

References

1. Snedecor GW, Cochran WG (1967) Statistical methods. Oxford and IBH, New Delhi
2. Wold S, Sjöström M, Eriksson L (2001) PLS-regression: a basic tool of chemometrics. Chemom Intell Lab Syst 58:109–130
3. Agresti A (1996) An introduction to categorical data analysis. Wiley, Hoboken
4. Everitt BS, Landau S, Leese M (2001) Cluster analysis, 4th edn. Arnold, London
5. Topliss JG, Costello RJ (1972) Chance correlation in structure-activity studies using multiple regression analysis. J Med Chem 15:1066–1068
6. Jaworska JS, Comber M, Auer C, Van Leeuwen CJ (2003) Summary of a workshop on regulatory acceptance of (Q)SARs for human health and environmental endpoints. Environ Health Perspect 111:1358–1360
7. Wold S (1978) Cross-validation estimation of the number of components in factor and principal components models. Technometrics 20:397–405
8. Roy K (2007) On some aspects of validation of predictive QSAR models. Expert Opin Drug Discov 2:1567–1577
9. Gramatica P (2007) Principles of QSAR models validation: internal and external. QSAR Comb Sci 26:694–701
10. Roy K, Mitra I (2011) On various metrics used for validation of predictive QSAR models with applications in virtual screening and focused library design. Comb Chem High Throughput Screen 14:450–474
11. Hawkins DM, Basak SC, Mills D (2003) Assessing model fit, by cross-validation. J Chem Inf Comput Sci 43:579–586
12. Roy K, Mitra I, Kar S, Ojha PK, Das RN, Kabir H (2012) Comparative studies on some metrics for external validation of QSPR models. J Chem Inf Model 52:396–408
13. Roy K, Chakraborty P, Mitra I, Ojha PK, Kar S, Das RN (2013) Some case studies on application of "r_m^2" metrics for judging quality of QSAR predictions: emphasis on scaling of response data. J Comput Chem 34:1071–1082
14. Mitra I, Roy PP, Kar S, Ojha P, Roy K (2010) On further application of rm2 as a metric for validation of QSAR models. J Chemometrics 24:22–33
15. Mitra I, Saha A, Roy K (2010) Exploring quantitative structure-activity relationship (QSAR) studies of antioxidant phenolic compounds obtained from traditional Chinese medicinal plants. Mol Simult 36:1067–1079
16. Golbraikh A, Tropsha A (2002) Beware of q2! J Mol Graph Model 20:269–276
17. Schuurmann G, Ebert RU, Chen J, Wang B, Kuhne R (2008) External validation and prediction employing the predictive squared correlation coefficient-Test-set activity mean vs training set activity mean. J Chem Inf Model 48:2140–2145
18. Consonni V, Ballabio D, Todeschini R (2010) Evaluation of model predictive ability by external validation techniques. J Chemometrics 24:194–201
19. Chirico N, Gramatica P (2011) 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 51:2320–2335

20. Roy K, Kar S (2014) How to judge predictive quality of classification and regression based QSAR models? In: Haq Z, Madura JD (eds) Frontiers in computational chemistry. Bentham Science Publishers, Sharjah
21. Fawcett T (2006) An introduction to ROC analysis. Pattern Recognit Lett 27:861–874
22. Perez-Garrido A, Helguera AM, Borges F, Cordeiro MNDS, Rivero V, Escudero AG (2011) Two new parameters based on distances in a receiver operating characteristic chart for the selection of classification models. J Chem Inf Model 51:2746–2759
23. Galvez J, Garcia-Domenech R, de Gregorio Alapont C, De Julian-Ortiz V, Popa L (1996) Pharmacological distribution diagrams: a tool for de novo drug design. J Mol Graph 14:272–276

http://www.springer.com/978-3-319-17280-4

A Primer on QSAR/QSPR Modeling
Fundamental Concepts
Roy, K.; Kar, S.; Das, R.N.
2015, X, 121 p. 47 illus., Softcover
ISBN: 978-3-319-17280-4