CHAPTER – IV
CROSS VALIDATION OF QSAR MODELS
Chapter IV

Cross Validation of QSAR Models

Cross validation in QSAR

QSAR studies started about 40 years ago. With the start of QSAR studies the quantitative descriptors were in the foreground. It was believed that the theoretical prediction did not play any vital role in development of new medicinal entities, it only play a minor role. A few physicochemical parameters, i.e. lipophilicity (log P), electronic properties, (expressed by σ), molar refractivity (MR), were used in the correlations. Later on along with quantum chemical and geometrical parameters: connectivity values, electrotopological state parameters, and many others features were tested, either they are suited to explain SARs in a quantitative manner and either the resulting models are capable to predict the activities of new analogs.

QSAR modeling is used to generate predictive models derived by applying statistical tools, correlating biological activity (such as required therapeutic effect and side effects) or physico-chemical properties in QSPR models for chemicals (drugs/toxicants/environmental pollutants) with descriptors representing the molecular structure and/or properties. These techniques are being applied in many disciplines for example risk assessment, toxicity prediction, and authoritarian decisions in addition to drug discovery and lead compound optimization. To obtain a good quality QSAR model dependence shifts on many factors, such as the quality of input data, the chosen descriptors and adopted or used statistical methods for modeling and for validation. Any QSAR modeling should ultimately lead to statistically robust and predictive models capable of making accurate and reliable predictions for the studied set of compounds or new compounds.

As a consequence of so many parameters and despite the fact that the use of too many parameters in one model has been criticized thirty years ago, thousands of meaningless chance correlations have been reported.
Rules, Procedure and conditions have been formed for QSAR studies to achieve valid correlations, meaningful variables shall be selected to develop the QSAR models; the significance of the correlation and each individual term in the regression model shall be justified by the appropriate statistical parameters. For the purpose, principle of parsimony shall be applied, i.e. results can be more or less equal, the simplest model shall be chosen from the generated models. In the process of generation of models use of too many variables shall be avoided, it is also believed that the tested model does not includes too many variables. In addition to these recommendation for cross-validation, Y scrambling, and external (test set) predictivity are used as validation criteria.

General validation principles used for quantitative structure-activity relationship (QSAR) models in the context of chemical regulation were developed due to the importance and implication of these methods in drug design. Later on, a brief analysis of different techniques used in validation of multiple linear regression models is reviewed and the hierarchical steps for the models validation are highlighted and a validation procedure is developed.

In the development process of model validation following statistical approaches are considered: correlation analysis (Pearson, Spearman and Kendall coefficients as parameters and associated significance levels), regression analysis (leave-one-out cross-validation and determination coefficients), and other inferential statistics (cross correlation coefficients, training vs. test experiment, correlated correlations analysis). The proposed statistical validation technique is privileged to ignore a QSAR model obtained by applying the molecular descriptors of similar family for the structure-activity relationship approach.

For validation of QSAR models usually following three strategies are adopted:

1. Internal validation or cross-validation;
2. External validation by splitting the available data set into training set for model development and prediction set for model predictivity check;
3. Blind external validation by application of model on new external data and
the success of any QSAR model depends on accuracy and type of the input data, selection of
appropriate descriptors and statistical tools, and most importantly validation of the developed
model. Validation is the process by which the reliability and relevance of a procedure or
developed model can be established for a specific purpose. For QSAR models validation must be
mainly for robustness, prediction performances and applicability domain of the models. Leave
one-out cross-validation generally leads to an over estimation of predictive capacity, and even
with external validation, no one can be sure whether the selection of training and test sets was
manipulated to maximize the predictive capacity of the model being published. Different aspects
of validation of QSAR models that need attention includes methods of selection of training set
compounds, setting training set size and impact of variable selection for training set models for
determining the quality of prediction. Development of novel validation parameters for judging
quality of QSAR models is also important.

Thus it becomes the responsibility of researcher who is developing the models to validate the
models and in extending the responsibility we have use the various cross-validation methods viz
internal validation and external validation.

Actually method of cross-validation evaluates the validity of a model by how well it predicts
data rather than how well it fits data.
Result and Discussion

In the cross-validation we have tested variety of cross validation parameters and recorded in the Analysis of Variance Table for the QSAR models or equations developed in chapter three.

Analysis Of Variance Table

<table>
<thead>
<tr>
<th>S. No.</th>
<th>K</th>
<th>N</th>
<th>PRESS</th>
<th>SSY</th>
<th>$R^2_{CV}$</th>
<th>PSE</th>
<th>$S_{PRESS}$</th>
<th>Press/SSY</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>95</td>
<td>69.29970</td>
<td>57.97960</td>
<td>0.19524</td>
<td>0.85409</td>
<td>0.86323</td>
<td>1.19524</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>95</td>
<td>51.22260</td>
<td>76.05670</td>
<td>0.32652</td>
<td>0.73429</td>
<td>0.74617</td>
<td>0.67348</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>95</td>
<td>40.40630</td>
<td>86.87310</td>
<td>0.53488</td>
<td>0.65217</td>
<td>0.66635</td>
<td>0.46512</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>95</td>
<td>32.35250</td>
<td>94.92680</td>
<td>0.65918</td>
<td>0.58357</td>
<td>0.59956</td>
<td>0.34082</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>95</td>
<td>28.30650</td>
<td>98.97290</td>
<td>0.71400</td>
<td>0.54586</td>
<td>0.56396</td>
<td>0.28600</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>94</td>
<td>27.20750</td>
<td>97.78580</td>
<td>0.72176</td>
<td>0.53800</td>
<td>0.55604</td>
<td>0.27824</td>
</tr>
<tr>
<td>7</td>
<td>5</td>
<td>93</td>
<td>25.87380</td>
<td>96.57380</td>
<td>0.73208</td>
<td>0.52746</td>
<td>0.54534</td>
<td>0.26792</td>
</tr>
<tr>
<td>8</td>
<td>5</td>
<td>92</td>
<td>24.73290</td>
<td>94.01640</td>
<td>0.73693</td>
<td>0.51849</td>
<td>0.53628</td>
<td>0.26307</td>
</tr>
<tr>
<td>9</td>
<td>5</td>
<td>91</td>
<td>23.33890</td>
<td>93.83440</td>
<td>0.75128</td>
<td>0.50643</td>
<td>0.52400</td>
<td>0.24872</td>
</tr>
<tr>
<td>10</td>
<td>5</td>
<td>90</td>
<td>22.59920</td>
<td>93.93020</td>
<td>0.75940</td>
<td>0.50110</td>
<td>0.51869</td>
<td>0.24060</td>
</tr>
<tr>
<td>11</td>
<td>5</td>
<td>89</td>
<td>21.90200</td>
<td>94.09770</td>
<td>0.76724</td>
<td>0.49607</td>
<td>0.51369</td>
<td>0.23276</td>
</tr>
</tbody>
</table>

Where K is the no. of parameters used in developed model.

N is the no of compounds on which model is tested.

PRESS (predicted residual sum squares).

SSY (sum of the squares of response value).

$r^2_{CV}$ (overall predictive ability).

PSE Predictive Squares Errors.

$S_{PRESS}$ (uncertainty of prediction).

For acquiring the validation of models, cross-validated parameters tested are PRESS (predicted residual sum squares), SSY (sum of the squares of response value), $r^2_{CV}$ (overall predictive ability), $S_{PRESS}$ (uncertainty of prediction).
For the uni-parametric model or first model the value of PRESS (predicted residual sum of squares) is higher than SSY (sum of the squares of response value) indicate that the model predicts just by chance and can't be considered statistically significant.

Furthermore, the ratio PRESS/SSY is used to estimate the confidence interval of the activity and explore the reasonability of the model.

In the case of first or uni-parametric model the ratio (PRESS/SSY) 1.19524 indicating that the proposed uni-parametric models is highly unreasonable QSAR models.

The indication of the performance of the model is obtained from $r^2_{cv}$ (the overall predictive ability). In this case the value of $r^2_{cv}$ is found very low and can be considered as statistically insignificant model and it has poor predictive power.

Another useful cross-validated parameter is $S_{PRESS}$, which is used in deciding uncertainty of prediction. However, this parameter in the present case is of no value as it is equivalent to the standard error of estimation (Se). Under such situation the parameter PSE is used. The lower value of PSE 0.85409 for this model indicates that the model has some correlation ability.

For the bi-parametric model the value of PRESS (predicted residual sum of squares) is lower than SSY (sum of the squares of response value) indicates that the model predicts better than chance and may be considered statistically significant.

The ratio of PRESS/SSY for the bi-parametric model representing the confidence interval 0.67348 of the activity and explore the reasonability of the model.

The further indication of the performance of the bi-parametric model is obtained from $r^2_{cv}$ (the overall predictive ability). In this case the value of $r^2_{cv}$ is found little higher than the uni-parametric model but can't be considered as statistically significant model and it has poor predictive power.
Another useful cross-validated parameter is $S_{\text{PRESS}}$, which is used in deciding uncertainty of prediction. However, this parameter in the present case is also of no value as it is equivalent to the standard error of estimation (Se). Thus the parameter PSE is used. The lower value of PSE 0.73429 for this model indicates that the model has some correlation ability.

The value of cross validation parameter PRESS (predicted residual sum of squares) for the tri-parametric model validation parameter as it is a good estimate of the real predictive error of the models. Its value less than SSY (sum of the squares of response value) indicates that the model predicts better than chance and can be considered statistically significant.

The ratio of PRESS/SSY for the tri-parametric model representing the confidence interval 0.46513 of the activity and explore the reasonability of the model.

The further indication of the performance of the tri-parametric model is obtained from $r^2_{cv}$ (the overall predictive ability). In this case the value of $r^2_{cv}$ is found little higher (0.54388) than the both uni and bi-parametric model and considered as statistically significant model but it has poor predictive power.

Also in this case $S_{\text{PRESS}}$, which is used in deciding uncertainty of prediction, is equivalent to the standard error of estimation (Se). Thus the parameter PSE is used. The lower value of PSE 0.65217 for this model indicates that the model has some correlation ability.

The value of cross validation parameter PRESS (predicted residual sum of squares) for the tetra-parametric model validation parameter as it is a good estimate of the real predictive error of the models. Its value (32.35250) less than SSY (sum of the squares of response value) (94.92680) indicates that the model predicts better than chance and can be considered statistically significant.

The ratio of PRESS/SSY for the tetra-parametric model representing the confidence interval 0.34082 of the activity and explore the reasonability of the model.
The further indication of the performance of the tetra-parametric model is obtained from $r^2_{cv}$ (the overall predictive ability). In this case the value of $r^2_{cv}$ is found little higher (0.65918) than the previous models and can be considered as statistically significant model but it has little poor predictive power.

Again in this case also $S_{\text{PRESS}}$, which is used in deciding uncertainty of prediction, is equivalent to the standard error of estimation (Se). Thus the parameter PSE is used. The lower value of PSE 0.58357 for this model indicates that the model has some better correlation ability.

The value of cross validation parameter PRESS (predicted residual sum of squares) for the penta-parametric model validation parameter as it is a good estimate of the real predictive error of the models. Its value (28.30650) less than SSY (sum of the squares of response value) (98.97290) indicates that the model predicts better than chance and can be considered statistically significant.

The ratio of PRESS/SSY for the penta-parametric model representing the confidence interval 0.28600 of the activity and explore the reasonability of the model.

The further indication of the performance of the penta-parametric model is obtained from $r^2_{cv}$ (the overall predictive ability). In this case the value of $r^2_{cv}$ is found little higher (0.71400) than the previous models and can be considered as statistically significant model but it has little poor predictive power.

Again in this case also $S_{\text{PRESS}}$, which is used in deciding uncertainty of prediction, is equivalent to the standard error of estimation (Se). Thus the parameter PSE is used. The lower value of PSE 0.54586 for this model indicates that the model has some better correlation ability.

The value of cross validation parameter PRESS (predicted residual sum of squares) for the penta-parametric model for 94 compounds is (27.20750) less than SSY (sum of the squares of response value) (97.78580) indicates that the model predicts better than chance and can be considered statistically significant.
The ratio of PRESS/SSY for the penta-parametric model representing the confidence interval 0.27824 of the activity and explore the reasonability of the model.

The further indication of the performance of the penta-parametric model is obtained from $r^2_{cv}$ (the overall predictive ability). In this case the value of $r^2_{cv}$ is found little higher (0.72176) than the previous models and can be considered as statistically significant model but it has little poor predictive power.

Again in this case also $S_{PRESS}$, which is used in deciding uncertainty of prediction, is equivalent to the standard error of estimation (Se). Thus the parameter PSE is used. The lower value of PSE 0.53800 for this model indicates that the model has some better correlation ability.

The value of cross validation parameter PRESS (predicted residual sum of squares) for the penta-parametric model for 93 compounds is (25.87380) less than SSY (sum of the squares of response value) (96.57380) indicates that the model predicts better than chance and can be considered statistically significant.

The ratio of PRESS/SSY for the penta-parametric model representing the confidence interval 0.26792 of the activity and explore the reasonability of the model.

The further indication of the performance of the penta-parametric model is obtained from $r^2_{cv}$ (the overall predictive ability). In this case the value of $r^2_{cv}$ is found little higher (0.73208) than the previous models and can be considered as statistically significant model but it has little poor predictive power.

Again in this case also $S_{PRESS}$, which is used in deciding uncertainty of prediction, is equivalent to the standard error of estimation (Se). Thus the parameter PSE is used. The lower value of PSE 0.52746 for this model indicates that the model has some better correlation ability.

The value of cross validation parameter PRESS (predicted residual sum of squares) for the penta-parametric model for 92 compounds is (24.73290) less than SSY (sum of the squares of response...
value) (94.01640) indicates that the model predicts better than chance and can be considered statistically significant.

The ratio of PRESS/SSY for the penta-parametric model representing the confidence interval 0.26307 of the activity and explore the reasonability of the model.

The further indication of the performance of the penta-parametric model is obtained from $r^2_{cv}$ (the overall predictive ability). In this case the value of $r^2_{cv}$ is found little higher (0.73693) than the previous models and can be considered as statistically significant model and it has significant predictive power.

Again in this case also $S_{PRESS}$, which is used in deciding uncertainty of prediction, is equivalent to the standard error of estimation (Se). Thus the parameter PSE is used. The lower value of PSE 0.51849 for this model indicates that the model has some better correlation ability.

The value of cross validation parameter PRESS (predicted residual sum of squares) for the penta-parametric model for 91 compounds is (23.33890) less than SSY (sum of the squares of response value) (93.8344) indicates that the model predicts better than chance and can be considered statistically significant.

The ratio of PRESS/SSY for the penta-parametric model representing the confidence interval 0.24872 of the activity and explore the reasonability of the model.

The further indication of the performance of the penta-parametric model is obtained from $r^2_{cv}$ (the overall predictive ability). In this case the value of $r^2_{cv}$ is found little higher (0.75128) than the previous models and can be considered as statistically significant model and it has significant predictive power.

Again in this case also $S_{PRESS}$, which is used in deciding uncertainty of prediction, is equivalent to the standard error of estimation (Se). Thus the parameter PSE is used. The lower value of PSE 0.50643 for this model indicates that the model has some better correlation ability.
The value of cross validation parameter PRESS (predicted residual sum of squares) for the penta-parametric model for 90 compounds is (22.59920) less than SSY (sum of the squares of response value) (93.93020) indicates that the model predicts better than chance and can be considered statistically significant.

The ratio of PRESS/SSY for this penta-parametric model representing the confidence interval 0.24060 of the activity and explore the reasonability of the model.

The further indication of the performance of this penta-parametric model is obtained from $r^2_{cv}$ (the overall predictive ability). In this case the value of $r^2_{cv}$ is found little higher (0.75940) than the previous models and can be considered as statistically significant model and it has significant predictive power.

Again in this case also $S_{PRESS}$, which is used in deciding uncertainty of prediction, is equivalent to the standard error of estimation (Se). Thus the parameter PSE is used. The lower value of PSE 0.50110 for this model indicates that the model has some better correlation ability.

The value of cross validation parameter PRESS (predicted residual sum of squares) for the penta-parametric model for 89 compounds is (21.90200) less than SSY (sum of the squares of response value) (94.09770) indicates that the model predicts better than chance and can be considered statistically significant.

The ratio of PRESS/SSY for this penta-parametric model representing the confidence interval 0.23276 of the activity and explore the reasonability of the model.

The further indication of the performance of this penta-parametric model is obtained from $r^2_{cv}$ (the overall predictive ability). In this case the value of $r^2_{cv}$ is found little higher (0.76724) than the previous models and can be considered as statistically significant model and it has significant predictive power.
Again in this case also $S_{\text{PRESS}}$, which is used in deciding uncertainty of prediction, is equivalent to the standard error of estimation (Se). Thus the parameter PSE is used. The lower value of PSE 0.49607 for this model indicates that the model has some better correlation ability.

In the present study, all the proposed models have PRESS < SSY indicating them to be better than chance and statistically significant.

Equation five onwards models are the models generated after the outlier of various compounds from the tested set of compounds and it is observed from the table of variance that the initially value of SSY also decreases along with the value of PRESS but at the very same time decreases in the ratio of both PRESS/ SSY indicates the significance of models along with justifying the deletion of various compounds from the calculation.

Furthermore, the ratio PRESS/SSY is used to estimate the confidence interval of the activity of compounds. To have a reliable QSAR model, PRESS/SSY should be 0.4 and the value of the ratio smaller than 0.1 is indicative of a excellent model. In our case, the ratio PRESS/SSY ranges between 1.19524– 0.23276 indicating that except model 1 to 3 all the proposed models are reasonable QSAR models. At this stage it is also worthy to mention that as we pass from the model 1 to 11 the reasonability of models increases. It is justify the addition of parameters in models and also the deletion of compounds from prediction.

The indication of the performance of the model is obtained from $r^2_{cv}$ (the overall predictive ability). In our case highest $r^2_{cv}$ is found for the model expressed by equation (11) consists of 5 parameters and tested on 89 compounds, indicating that it has a good predictive power.

Another useful cross-validated parameter is $S_{\text{PRESS}}$, which is used in deciding uncertainty of prediction. However, this parameter in the present case is of no value as it is equivalent to the standard error of estimation (Se). Under such situation the parameter PSE is used. The lowest value of PSE, the better is the predictive power which indicates that the models have considerable correlation ability. Based on PSE values (Analysis of Variance Table) once again
we observed that the model 11 has the best correlation ability and predictability amongst all the models developed in present study.

Inter-relationship between the number of parameters and number of compounds with various cross validation parameters are presented in form of bar graph from figure 1 to 6.

Figure 1. Graphical representation of change in the values of PRESS from model 1 to 11.
Figure 2. Graphical representation of change in the values of SSY from model 1 to 11.
**Figure 3.** Graphical representation of change in the values of $R^2_{CV}$ from model 1 to 11.
Figure 4. Graphical representation of change in the values of PSE from model 1 to 11.
Figure 5. Graphical representation of change in the values of $S_{PRES}$ from model 1 to 11.
Figure 6. Graphical representation of change in the values of PRESS/SSY from model 1 to 11.