Chapter - 4

Results and Discussion

Essentially all models are wrong, but some are useful. The practical question is how wrong do they have to be to not be useful.

— George E. P. Box
4.1. **QSAR equation: Analysis, interpretation and extrapolation**

The receptor and ligand-based studies described in Chapters II and III involve the development (and validation) of QSAR models primarily based on interaction of the ligand molecule with individual receptor active site residues/amino acids and different probe atoms placed at intersection points of a three-dimensional lattice grid (representing a receptor surrogate) respectively. Various QSAR models (equations and their associated statistical parameters) that were developed for different datasets have already been mentioned in these chapters. We will now discuss how a QSAR equation generated based on these two types of studies looks like, what it contains, what it represents, its analysis, interpretation, and finally its use to predict the activity of a new molecule. These aspects can also be found in some of the related published work.\(^1\)\(^3\)

### 4.1.1. Receptor-based QSAR equation

Consider eq. 1 which actually is the receptor-based WSETH model developed using the Cerius2 program\(^4\) for glycogen phosphorylase b (GPB) inhibitors as shown in Table 2.08 in Chapter II:

\[
pK_i = 2.44 - 0.56 \text{ (E}_{\text{Ala383}}) - 1.15 \text{ (S}_{\text{Ala140}}) - 0.66 \text{ (T}_{\text{Asn376}}) + 0.75 \text{ (T}_{\text{Ala456}}) + 0.82 \text{ (Lipo)} + 0.37 \text{ (S}_{\text{Gly134}}) + 0.49 \text{ (H}_{\text{Gly675}})
\]

**eq. 1**

Eq. 1 is a mathematical correlation between the biological activity (pK\(_i\)) and the descriptors or dimensions employed. Different descriptors, here, include the *per-residue* terms (non-bonded or non-covalent interaction energies between the ligand and the active site residues in the receptor *i.e.*, steric, electrostatic, total non-bonded and hydrogen-bonding energies) and the *whole-molecule* terms (*e.g.*, Lipo, Coul, vdW, Hbond etc.) described in detail in Section 2.4 of Chapter II. So, in other words, eq. 1 depicts how these specific *per-residue* and *whole-molecule* terms are involved in modulating the molecular activities. While analyzing such a QSAR equation, one should remember that more negative the value of the interaction energies, stronger is the interaction between the ligand and the receptor. Similarly positive values of these interaction energies imply weaker interaction between the respective groups of the ligand and the receptor. However, it is the sign of the coefficient of these descriptors/terms in the QSAR equations that will ultimately decide whether to strengthen/increase (*i.e.*, make the interaction energy more negative) or weaken/decrease (*i.e.*, make the interaction energy relatively more positive) the interaction, in order to improve
binding. In case of descriptors like ‘Lipo’ (a term representing hydrophobicity), favourable interactions have positive values and unfavourable interactions are negative. The magnitude of the descriptors in the QSAR equation indicates the extent of their correlation with the biological activity. Thus the overall molecular activity can be increased by reducing the magnitude of those descriptors, which are negatively correlated with the activity (so as to change their ‘-‘ sign to ‘+’), and at the same time increasing the magnitude of those descriptors correlated positively with the activity (so as to intensify their ‘positive’ contributions).

The electrostatic (Coulombic) and steric (van der Waals) interactions of the ligand with receptor residues Ala383 and Ala140 respectively appear with negative coefficients in eq. 1. This indicates that the biological activity can be amplified by strengthening the steric and electrostatic interactions of the ligand with these residues of the enzyme. Similarly, enhancing the total non-bonded (steric and electrostatic) interaction of the ligand with Asn376 will increase the binding affinity, due to the negative coefficient of this interaction in the equation. On the other hand, because of the positive coefficients of the steric, H-bonding and total non-bonded interactions of the ligand with residues Gly134, Gly675, and Ala456 respectively, fading the vigour of the interactions of the ligand with these particular residues of the receptor will favour binding. However these interactions appear to be very sensitive and thus require cautious modifications so as to reduce their strength without adversely affecting the overall binding of the ligand with the receptor. Similarly the biological activity can also be augmented by rendering the ligand molecule more lipophilic, due to the positive coefficient of the whole-molecule hydrophobic term ‘Lipo’ in the equation. For predicting the biological activity of a new molecule with the QSAR model shown in eq.1, it is subjected to the same protocol i.e., energy minimization followed by docking into the respective receptor active site, and finally substituting the descriptors (per-residue and whole-molecule terms), thus obtained, into eq. 1 to obtain the predicted binding affinity of the molecule.

4.1.2. Ligand-based QSAR equation

Now consider eq. 2 which is the ligand-based WSETTH model developed using the Cerius2 program for the angiotensin converting enzyme (ACE) inhibitors as shown in Table 3.08 in Chapter III:

\[
\text{pIC}_{50} = 0.99 - 0.18 (T\_308) - 0.28 (S\_644) - 0.59 (H\_366) + 0.18 (S\_295) + 0.96 (H\_279) \\
0.22 (S\_414) - 0.15 (T\_494) + 0.49 (E\_15)
\]

eq. 2
An equation generated by a ligand-based QSAR model (like the one shown above) is analyzed and interpreted exactly in the same way as explained earlier for the receptor-based model. The only difference is that the descriptors/terms in the equation represent the interaction of the ligand with various probe atoms placed at the intersection points of a 3D-lattice grid (representing a receptor surrogate), and not with real receptor residues. The different interaction energy values (i.e., steric, electrostatic, total non-bonded and H-bonding energies) thus obtained as per-grid point descriptors, were employed along with the whole-molecule set of variables, to develop the QSAR model as described in detail in Chapter III.

According to eq. 2, an enhancement in the activity of the molecule can be achieved by strengthening its total non-bonded interactions with the probe atoms at grid points 308 and 494, due to the negative coefficients of these interactions in the equation. For similar reasons, an upsurge in the interactions of the ligand with the steric probes at grid points 414 and 644 will result in amplification of its biological activity. On the other hand, due to their positive coefficients in the equation, weakening of the interactions of the ligand with the electrostatic and steric probes at grid points 15 and 295 respectively can increase its activity. Similarly an increase in the interaction of the ligand with the H-bonding probe at grid point 366 will lead to an enhancement in the molecular activity due to its negative coefficient in the equation. However strengthening the ligand's interaction with the H-bonding probe at grid point 279 would adversely affect the activity due to its positive coefficient in the equation. The activity of a new ligand can be predicted with a ligand-based QSAR model like eq. 2 by aligning the energy-minimized molecule over the pre-superimposed training set of molecules of the respective dataset, placing it in the same orientation in the 3D-grid, computing it's interactions with different probe atoms placed at the intersection points of the grid, and finally substituting the per-grid point terms thus obtained into the QSAR equation along with the whole-molecule descriptors, to obtain the predicted activity of the molecule.

4.2. Frequency analysis of QSAR equations

Several QSAR models were developed by incorporating various per-residue or per-grid point terms (i.e., steric, electrostatic, total non-bonded interaction energy, and H-bonding terms) in different possible combinations, along with all the whole-molecule set of descriptors. After analyzing all combinations of models for each dataset, it was found that WSETH models outperformed the others in terms of the statistical parameters. This was observed both in the receptor as well as the ligand-based studies described in Chapters II and III respectively.
Hence for simplicity, only the WSETH models have been discussed comprehensively for all the datasets in this chapter. This is also justified by the fact that these models are based on the entire set of dimensions investigated in this thesis viz. steric (S), electrostatic (E), total non-bonded (T), and hydrogen-bonding (H) interactions along with the whole-molecule descriptors (W).

The descriptors that appeared in the best QSAR equations are not the only terms responsible for the molecular activity. They just represent one of the solutions to the problem that has been best fitted mathematically to explain the activity. There might be many other terms which are important in modulating the biological activity, but they died during the course of evolution (via G/PLS) and could not make it to the best models. One method of extracting information about these terms consists of analyzing the frequency at which each descriptor appears in the final population of equations. Hence the WSETH models of all the datasets were subjected to frequency analysis and given below are the frequency plots of the most repeatedly occurring descriptors for different datasets. The frequency of occurrence of different descriptors is shown on the x-axis, whereas the signs of the terms in the equations are shown on the y-axis in terms of the percentage (%). Terms with positive coefficients in the equations are displayed as ‘positive frequency’ values, whereas those with negative coefficients are shown with ‘negative frequencies’.

As mentioned earlier, in addition to the terms/descriptors, the signs (+ or -) of their coefficients in the QSAR equations are also equally important in regulating the biological activity of the molecules. Hence the contribution of each individual dimension towards the total pool of 500 QSAR equations of the WSETH models was also analyzed for every dataset and plotted as 2D-pie charts along with their distribution into terms with negative and positive coefficients as shown in Figures 4.01 to 4.06 for receptor-based studies and in Figures 4.07 to 4.12 for ligand-based studies. The overall trends of variation of different dimensions in the WSETH models across all the datasets have been summarized in 2D-line plots given in Figures 4.13 and 4.14 for the receptor and ligand-based studies respectively.
Figure 4.01a. Frequency plot of descriptors appearing repetitively in the WSETH models developed for the thrombin dataset. S, E, T, H refers to the steric, electrostatic, total non-bonded and H-bonding interactions of the ligand with the respective residues of the receptor active site.

Figure 4.01b. Distribution of different descriptors in the total pool of 500 equations of the WSETH model developed for the thrombin dataset (a), along with their further dissemination into terms with negative (b) and positive (c) coefficients in the same set of equations. S, E, T, H, and W refer to the steric, electrostatic, total non-bonded, hydrogen-bonding and the whole-molecule dimensions respectively.
Figure 4.02a. Frequency plot of descriptors appearing repetitively in the WSETH models developed for the glycogen phosphorylase b (GPB) dataset. S, E, H refers to the steric, electrostatic, and H-bonding interactions of the ligand with the respective residues of the receptor active site. The term 'Lipo' stands for a variable describing the hydrophobic character of the molecule.

Figure 4.02b. Distribution of different descriptors in the total pool of 500 equations of the WSETH model developed for the glycogen phosphorylase b (GPB) dataset (a), along with their further dissemination into terms with negative (b) and positive (c) coefficients in the same set of equations. S, E, T, H, and W refer to the steric, electrostatic, total non-bonded, hydrogen-bonding and the whole-molecule dimensions respectively.
Figure 4.03a. Frequency plot of descriptors appearing repetitively in the WSETH models developed for the thermolysin dataset. S, E, T, H refers to the steric, electrostatic, total non-bonded and H-bonding interactions of the ligand with the respective residues of the receptor active site.

Figure 4.03b. Distribution of different descriptors in the total pool of 500 equations of the WSETH model developed for the thermolysin dataset (a), along with their further dissemination into terms with negative (b) and positive (c) coefficients in the same set of equations. S, E, T, H, and W refer to the steric, electrostatic, total non-bonded, hydrogen-bonding and the whole-molecule dimensions respectively.
Figure 4.04a. Frequency plot of descriptors appearing repetitively in the WSETH models developed for the HIV-1 protease dataset. S, E, T refers to the steric, electrostatic, and total non-bonded interactions of the ligand with the respective residues of the receptor active site.

Figure 4.04b. Distribution of different descriptors in the total pool of 500 equations of the WSETH model developed for the HIV-1 protease dataset (a), along with their further dissemination into terms with negative (b) and positive (c) coefficients in the same set of equations. S, E, T, H, and W refer to the steric, electrostatic, total non-bonded, hydrogen-bonding and the whole-molecule dimensions respectively.
Figure 4.05a. Frequency plot of descriptors appearing repetitively in the WSETH models developed for the HIV-1 integrase dataset. S, E, T, H refers to the steric, electrostatic, total non-bonded and H-bonding interactions of the ligand with the respective residues of the receptor active site. CvdW is the modified Coulombic-van der Waals interaction energy, Emod is the model energy and Esite is a term describing polar interactions in the active site.

Figure 4.05b. Distribution of different descriptors in the total pool of 500 equations of the WSETH model developed for the HIV-1 integrase dataset (a), along with their further dissemination into terms with negative (b) and positive (c) coefficients in the same set of equations. S, E, T, H, and W refer to the steric, electrostatic, total non-bonded, hydrogen-bonding and the whole-molecule dimensions respectively.
Figure 4.06a. Frequency plot of descriptors appearing repetitively in the WSETH models developed for the HLA-binding peptide dataset. S, E, T, H refers to the steric, electrostatic, total non-bonded and H-bonding interactions of the ligand with the respective residues of the receptor active site.

Figure 4.06b. Distribution of different descriptors in the total pool of 500 equations of the WSETH model developed for the HLA-binding peptide dataset (a), along with their further dissemination into terms with negative (b) and positive (c) coefficients in the same set of equations. S, E, T, H, and W refer to the steric, electrostatic, total non-bonded, hydrogen-bonding and the whole-molecule dimensions respectively.
Figure 4.07a. Frequency plot of descriptors appearing repetitively in the WSETH models developed for the bradykinin-like peptide dataset. S, E, H refers to the steric, electrostatic, and H-bonding interactions of the ligand with the respective probe atoms at specific grid points. SASA is the solvent accessible surface area and vol stands for the molecular volume.

Figure 4.07b. Distribution of different descriptors in the total pool of 500 equations of the WSETH model developed for the bradykinin-like peptide dataset (a), along with their further dissemination into terms with negative (b) and positive (c) coefficients in the same set of equations. S, E, T, H, and W refer to the steric, electrostatic, total non-bonded, hydrogen-bonding and the whole-molecule dimensions respectively.
Figure 4.08a. Frequency plot of descriptors appearing repetitively in the WSETH models developed for the angiotensin converting enzyme (ACE) dataset. S, T, H refers to the steric, total non-bonded and H-bonding interactions of the ligand with the respective probe atoms at specific grid points.

Figure 4.08b. Distribution of different descriptors in the total pool of 500 equations of the WSETH model developed for the angiotensin converting enzyme (ACE) dataset (a), along with their further dissemination into terms with negative (b) and positive (c) coefficients in the same set of equations. S, E, T, H, and W refer to the steric, electrostatic, total non-bonded, hydrogen-bonding and the whole-molecule dimensions respectively.
Figure 4.09a. Frequency plot of descriptors appearing repetitively in the WSETH models developed for the benzodiazepine-receptor (BZR) agonist dataset. S, E, T, H refers to the steric, electrostatic, total non-bonded and H-bonding interactions of the ligand with the respective probe atoms at specific grid points. SASA is the solvent accessible surface area and vol stands for the molecular volume.

Figure 4.09b. Distribution of different descriptors in the total pool of 500 equations of the WSETH model developed for the benzodiazepine-receptor (BZR) agonist dataset (a), along with their further dissemination into terms with negative (b) and positive (c) coefficients in the same set of equations. S, E, T, H, and W refer to the steric, electrostatic, total non-bonded, hydrogen-bonding and the whole-molecule dimensions respectively.
Figure 4.10a. Frequency plot of descriptors appearing repetitively in the WS3TH models developed for the acetylcholinesterase (AchE) dataset. S, E, T, H refers to the steric, electrostatic, total non-bonded and H-bonding interactions of the ligand with the respective probe atoms at specific grid points.

Figure 4.10b. Distribution of different descriptors in the total pool of 500 equations of the WS3TH model developed for the acetylcholinesterase (AchE) dataset (a), along with their further dissemination into terms with negative (b) and positive (c) coefficients in the same set of equations. S, E, T, H, and W refer to the steric, electrostatic, total non-bonded, hydrogen-bonding and the whole-molecule dimensions respectively.
Figure 4.11a. Frequency plot of descriptors appearing repetitively in the WSETH models developed for the dihydrofolate reductase (DHFR) dataset. S, E, T, H refers to the steric, electrostatic, total non-bonded and H-bonding interactions of the ligand with the respective probe atoms at specific grid points.

Figure 4.11b. Distribution of different descriptors in the total pool of 500 equations of the WSETH model developed for the dihydrofolate reductase (DHFR) dataset (a), along with their further dissemination into terms with negative (b) and positive (c) coefficients in the same set of equations. S, E, T, H, and W refer to the steric, electrostatic, total non-bonded, hydrogen-bonding and the \textit{whole-molecule} dimensions respectively.
Figure 4.12a. Frequency plot of descriptors appearing repetitively in the WSETH models developed for the cyclooxygenase-2 (COX-2) dataset. S, E, T, H refers to the steric, electrostatic, total non-bonded and H-bonding interactions of the ligand with the respective probe atoms at specific grid points.

Figure 4.12b. Distribution of different descriptors in the total pool of 500 equations of the WSETH model developed for the cyclooxygenase-2 (COX-2) dataset (a), along with their further dissemination into terms with negative (b) and positive (c) coefficients in the same set of equations. S, E, T, H, and W refer to the steric, electrostatic, total non-bonded, hydrogen-bonding and the whole-molecule dimensions respectively.
Figure 4.13. 2D-line plots depicting distribution trends of different descriptors in the total pool of 500 equations of the WSETH models developed for all the receptor-based datasets (a), along with their further dissemination into terms with negative (b) and positive (c) coefficients in the same set of equations. S, E, T, H, and W refer to the steric, electrostatic, total non-bonded, hydrogen-bonding and the whole-molecule dimensions respectively.
Figure 4.14. 2D-line plots depicting distribution trends of different descriptors in the total pool of 500 equations of the WSETH models developed for all the ligand-based datasets (a), along with their further dissemination into terms with negative (b) and positive (c) coefficients in the same set of equations. S, E, T, H, and W refer to the steric, electrostatic, total non-bonded, hydrogen-bonding and the whole-molecule dimensions respectively.
A careful examination of the distribution graphs (shown as pie charts and line plots above) revealed that steric and electrostatic terms are the most frequently occurring and thus have the highest contribution to the total population of WSETH equations in receptor-based studies, followed by the total non-bonded and H-bonding terms. Whole-molecule descriptors are found to have the least contribution to the total pool of equations. However in ligand-based studies, H-bonding and steric descriptors constitute the bulk of the final population of equations, followed by the total non-bonded and electrostatic terms. The whole-molecule variables were the least frequent in these studies too. Another thing which can be observed from the distribution pattern of different descriptors is that in most of the receptor or ligand-based datasets, the interaction energy terms (S, E, T, and H) with negative coefficients (i.e., which are negatively correlated with the activity) in the QSAR equations have a significant edge in terms of number over those with the positive coefficients. This is indicative of the fact that the terms recommending strengthening of the interactions between the ligand and specific receptor residues are in significant majority than those suggesting dwindling of some of these interactions in order to have an overall improvement in the biological activities of the molecules. In simpler terms it accentuates that in most of the molecules there is ample scope to improve the activity. The equation generated for any receptor or ligand-based QSAR model can easily be analyzed, interpreted and applied for prediction as explained comprehensively in Section 4.1. It can also be displayed graphically in a simplified manner as shown in Figures 4.15 and 4.16 for the WSETH models of HIV-1 integrase (receptor-based) and ACE (ligand-based) inhibitors respectively. In Figure 4.15, the ligand-residue interactions shown as green arrows need to be enhanced (due to their negative coefficients in the QSAR equations) and those displayed as red arrows are required to be diminished (due to their positive coefficients in the QSAR equations) to augment the molecular activity. Similarly in Figure 4.16, green spheres (e.g., S_414) represent areas where specific interactions of the ligand (i.e., steric) with probe atoms at particular grid points (i.e., 414) needs to be strengthened (due to their negative coefficients in the QSAR equations) whereas red spheres (e.g., H_279) indicate areas where specific interactions of the ligand (i.e., H-bonding) with probe atoms at particular grid points (i.e., 279) needs to be weakened (due to their positive coefficients in the QSAR equations) to enhance the activity of the ligand molecule.
Figure 4.15. Graphical representation of the WSETH model of HIV-1 integrase inhibitors, depicting the interactions of the ligand molecule with the important active site residues.

Figure 4.16. Graphical representation of the WSETH model of ACE inhibitors, showing important regions surrounding ligand molecule where specific modifications can be done to improve the activity.

The statistical indicators of all the QSAR models developed for the receptor and ligand-based datasets are reported in Chapters II and III respectively. Different models constructed were found to be reasonably acceptable in terms of most of these parameters. After randomization\textsuperscript{5} of the activity data, the $r^2$ values decreased to smaller numbers, negating the possibility of a chance correlation. Cross-validation\textsuperscript{6} by leave-one out and leave-five-out procedures returned statistically significant $q^2$ values. The bootstrap\textsuperscript{7} results further advocated the robustness of the models. The predictive $r^2$ of all the models on the test set was also satisfactory enough, indicating good predictive and extrapolative power of the models for molecules outside the training set. Let's have a look at the trends of variation of each of these statistical parameters of the various QSAR models developed for all the receptor-based (Figure 4.17, a1 - j1) and ligand-based (Figure 4.17, a2 - j2) datasets described in Chapters II and III respectively. In Figure 4.17, the y-axis shows the observed range of a particular statistical parameter, whereas the x-axis represents the 15-QSAR models that were developed for all datasets using different combinations of dimensions as explained in detail in Chapters II and III for receptor and ligand-based studies respectively.
Chapter Four
Figure 4.17. 2D-line plots representing the trends of variation of different statistical parameters of the WSETH models developed for the receptor-based (a1 - j1) and ligand-based (a2 - j2) datasets

N = Number of components
L = Length (number of terms) of the QSAR equation including the constant
r2 = Conventional (non cross-validated) correlation coefficient (r²)
PRESS = Predicted residual sum of squares
r2bs and sdsbs = Mean values of correlation coefficient and standard deviation respectively from bootstrap analysis (r²bs and sdsbs)
r2random = Mean value of the correlation coefficient after randomization of the dependent variable at 99% confidence interval (r²random)
q2(L1O) and q2(L5O) = Cross-validated correlation coefficient from leave-one-out and leave-five-out respectively [q²(L-1-O) and q²(L-5-O)]
r2pred = Predictive correlation coefficient of the test set (r²pred)
4.4. Influence of multidimensionality on statistical parameters

The influence of increasing dimensions on different statistical parameters of the receptor and ligand-based QSAR models described in Chapters II and III respectively is given below:

4.4.1. Effect on the number of components (N)

As seen in Figure 4.17 (a1 and a2), the total number of components increased as more and more dimensions (descriptors) were included in the model development. However this was also followed by significant enhancement in the computational time to derive the QSAR model, in addition to the introduction of noise/error or reduction in the signal to noise ratio. This was probably due to a rise in the number of solutions (owing to the increasing number of components) made available to the genetic algorithm for cross-overs and mutations. A precipitous enhancement in the $r^2$ values of the models was observed with an initial increase in the number of components particularly from 4 to 6, which later kept on fading with further increase in the components. The optimum number of components at which most of the QSAR models had good correlation statistics with minimal errors was observed to be 6 to 7.

4.4.2. Effect on the length of equations (L)

As can be concluded from Figure 4.17 (b1 and b2), the total number of terms in the equations (length of equations) required to develop the best QSAR models was found to increase as more number of dimensions were added. Like in case of the number of components (N), increasing the length of the QSAR equations resulted in an upsurge in the computational time which compounded significantly when clubbed with enhanced number of components. Taking into consideration the computational time and other parameters, the optimum number of terms which produced statistically best QSAR equations in most of the receptor and ligand-based studies was found to be 7.

4.4.3. Effect on the correlation coefficients

As perceived from Figure 4.17 (c1 and c2), the correlation coefficient ($r^2$) was observed to increase upon incremental incorporation of additional dimensions. The amplification in $r^2$ values was more pronounced at higher number of components and increased number of terms in the QSAR equations. Also as highlighted in Figure 4.17, the other related cross-validation$^6$ parameters like $r^2_{bs}$ (e1 and e2), $q^2$ (L-1-O and L-5-O in Figures h1, h2 and i1, i2 respectively), as well as $r^2_{pred}$ were also found to respond in a similar manner as the $r^2$ values.
with respect to the addition of more dimensions. Most of the satisfactory QSAR models were observed to have their correlation coefficient values in the range of 0.75 to 0.80.

4.4.4. Effect on the error parameters

With incorporation of increasing number of dimensions in the models, the predicted residual sum of squares (PRESS) values were found to be reduced, however the reduction was not to a large extent as can be seen in Figure 4.17 (d1 and d2). The receptor-based studies were observed to have larger PRESS values (average 60) compared to the ligand-based studies (average 45). This might probably be due to the involvement of too many parameters in the former, relative to the later types of studies. In case of the standard deviation parameter arising from the bootstrap analysis (sdbs), no general trend was observed as is apparent in Figure 4.17 (f1 and f2). The values of sdbs were found to vary in the range of 0.02 – 0.14 and 0.05 – 0.30 for the receptor and ligand-based studies respectively.

4.4.5. Effect on the randomization procedure

After scrambling or shuffling the activity values throughout each dataset, the $r^2$ values were observed to fall drastically. As can be realized from Figure 4.17 (g1 and g2), no general trend or variation could be found from this observation. The average value of $r^2$ after randomization or y-scrambling test (i.e., $r^2_{\text{random}}$) was observed to be 0.31 and 0.32 for the receptor and ligand-based studies respectively.

4.5. General deductions from the studies

In addition to the aforementioned specific results, the overall work has also highlighted several significant observations from which following conclusions could be derived:

- No structural dimension has been found to be universally acceptable. The one that is predictive for a particular data set may be no better than existing approaches for a typical QSAR data set. So one should actually attempt every possible, but strictly relevant, dimension that best suits his/her dataset of interest.

- In most of the receptor and ligand-based studies described in Chapters II and III respectively, the per-residue or per-grid point dimensions like steric and electrostatic have been found to be significantly dominating the final outcome of the QSAR model. The whole-molecule dimensions appeared only occasionally in the best models; and amongst them those representing hydrophobicity (i.e., Lipo and logP) have shown the maximum frequency. This
indicates that van der Waals, Coulombic and hydrophobic interactions (to an extent) are the major driving forces responsible for protein-ligand binding with respect to the datasets being investigated.

- The position of the grid box considerably influenced the statistics particularly the number of components in the ligand-based (or alignment-based) QSAR models. Thus in order to examine the influence of the overall orientation of the aligned molecules in the grid, the entire aggregate of the overlaid molecules was rotated systematically in steps of 90° around the x, y and z-axes, and the optimum number of components and \( q^2 \) values were obtained for each orientation. The orientation of the aligned molecules with the best \( q^2 \) value was selected for further model development.

- Employing a significantly large number of dimensions has been observed to result in chance correlation. Selecting 5 – 6 data-points per variable, circumvents the possibility of such correlation for majority of the intermediate size datasets used in the study.

- Incorporating numerous correlated or illogical/irrational dimensions has been found to produce models which have good internal statistics, but they generally reduce the signal-to-noise ratio and are unable to accurately and consistently explain the behaviour of molecules outside the training sets.

- Including a wide range of different, meaningful and essentially independent/non-correlated dimensions, biophysically related to the property being predicted, may not always enhance the internal or external predictivity of a QSAR model, but they certainly potentiate the interpretability and general applicability of the model.

- Models generated with raw or original values of the descriptors were found to be significantly inferior in terms of statistical parameters compared to the models developed with descriptors scaled to zero mean and unit standard deviation. This demonstrated the importance of standardizing descriptors by assigning equal weightage and putting them all on the same platform for a meaningful statistical analysis.

- Models built with an optimally large and structurally/biologically diverse subset of the whole dataset tend to be less noisy than those developed with the entire compiled dataset. For example in most of the datasets (more precisely COX-2, DHFR, BZR, AchE, HLA), the number of molecules compiled from the literature were much more than those actually
employed. Redundant or similar molecules were dropped and only those with significant
diversity were selected from the entire dataset and used for model development. Care was
taken to include all possible/available variations in the molecules without introducing much
redundancy. This mapping of the substituent (descriptor) space with minimum number of
compounds substantially reduced the noise in the models generated and improved their
statistical performance. However in cases like bradykinin and ACE peptides where the
number of molecules available in the literature were less, the entire datasets were employed
for model development.

4.6. References

   comprehensive analysis of the thermodynamic events involved in ligand-receptor binding using CoRIA and its
   Model.* 2007, 47, 2345-2357.
6. Richard, D.; Cramer III, R. D.; Bunce, J. D.; Patterson, D. E.; Frank, I. E., Crossvalidation,
   Bootstrapping, and Partial Least Squares Compared with Multiple Regression in Conventional QSAR Studies.