Chapter 3

QSAR Methodology and ADAPT

The previous chapter described a number of modeling techniques with which QSAR models can be built. Based on the nature of the method used, QSAR models are classified as linear or nonlinear. However, the modeling process does not simply consist of passing data through an algorithm. As described in Chapter 1, the fact that we cannot directly calculate physical properties or biological activities requires us to take an indirect route. As a result, QSAR modeling is a stepwise process consisting of five main steps:

1. Structure entry and optimization
2. Descriptor calculations
3. Objective and subjective feature selection
4. Model development
5. Prediction

This is a very broad overview and certain steps, such as set generation and interpretation have been skipped over, though these will be discussed in more detail in subsequent chapters. Another important step in the QSAR model development process is the consideration of the validity of models. This aspect has many facets and one of them involves deciding as to whether a model will be applicable to a set of unseen query compounds. This topic is discussed in more detail in Chapter 5.

The utility of a QSAR model depends on its intended future use. If a model is to be used as a screen in a high throughput pipeline, the predictive ability of the model is paramount. In these cases the high predictive ability of CNN's and random forests make models based on these methods attractive. If a QSAR model is to be used as a guide to possible modifications of molecules to improve their activities, the interpretability of the model assumes a major role. In this case the simple PLS interpretation scheme that can be applied to linear models make them good candidates for QSAR modeling inspite of their lower predictive performance for biological properties. Chapters 8 and 9 describe two approaches to providing interpretations for CNN models. As a result, one gets the best of both worlds - high predictive ability and a measure of interpretability.

This chapter discusses in detail the various steps in the model development process and is oriented towards the use of the ADAPT[1,2] software package for 2-D QSAR model development and testing.

3.1 Structure Entry and Optimization

The model building process using ADAPT begins with the creation of a work area which initializes the various files and related storage requirements for a QSAR study. The molecules to be used in the study can be available as 2-D or 3-D structures. In general, the data are in the former format and as a result, 3-D structures are required. 3-D structures are usually generated using Hyperchem. The resultant structures are crudely optimized using a molecular mechanics method within Hyperchem. Once the dataset has been converted to 3-D structures, they are rigorously optimized with Mopac 7.01. This program employs a semi-empirical method using the PM3[3,4] Hamiltonian. This Hamiltonian is reported to be well suited to the purpose of geometry optimization. Since some molecular descriptors also require information about the electronic environment of the molecule, the molecules are also optimized for electronic properties. In this case the AM1[5–7] Hamiltonian is used. Once the molecules have been optimized for geometry and electronics, they are stored in the ADAPT work area.

3.2 Molecular Descriptor Calculations

As mentioned before, the fundamental assumption of QSAR modeling is that molecular structure can be correlated to physical or biological properties. Thus the fundamental requirement is some method to encode various structural features in a molecule. Molecular descriptors fulfill this requirement. Descriptors are (in general) numerical representations of specific molecular features. Such features can range from very simple ones such as the number of carbons or number of halogen atoms to more complex and abstract features such as graph invariants of the molecular graph or the information content of a molecule as characterized by entropy. Several packages are available to calculate a wide variety of descriptors. Examples include Dragon,[8] JOELib[9] and ADAPT. Owing to the large variety of descriptors that can be calculated we restrict ourselves to a discussion of the main types of descriptors that are calculated by ADAPT and refer the reader to the literature[10] for additional information. The descriptors calculated by ADAPT can be classified into four types: geometrical, topological, electronic and hybrid. The following sections describe the nature of each descriptor class in detail.

3.2.1 Geometric Descriptors

Geometric descriptors characterize the shape and extent of the molecule in terms of its 3-D coordinates. As a result accurate coordinates are required and so the structure must be geometry optimized before these descriptors can be calculated. Examples include moment of inertia,[11] molecular surface area and volumes,[12] and shadow descriptors.[13,14] The surface area and volume descriptors are usually used in combination with atomic properties (such as partial charges or hydrophobicities) and are useful in characterizing the distribution of these properties. The shadow area descriptors align the first two moments of inertia of the molecule along the X and Y axes and then calculate the area of the projection of the molecule on the XY, XZ and YZ planes. In general these types of descriptors capture features related to molecular size and shape and thus are generally physically interpretable. The drawback to these descriptors is that they require accurate molecular geometries and thus for large sets of molecules the optimization step can become time consuming. Furthermore, the ADAPT implementation of these descriptors do not take conformational features into account and work with the lowest energy conformer. Situations where conformation details play an important role (such as ligand binding) will not be accurately characterized by these descriptors.

3.2.2 Topological

As the name suggests, topological descriptors consider the topology of a molecule. That is, in the most general case, only the connections between the atoms in a hydrogen suppressed molecule, effectively converting it into a mathematical graph. Certain topological descriptors consider the type or certain properties of atoms involved in the connections as weights. Topological descriptors characterize features such as path lengths and connectivity. Examples include connectivity indices,[15–17] distance edge vectors[18] and eccentricity indices.[19] Since topological descriptors consider the molecule as a mathematical graph, a number of these descriptors are simply various graph invariants or other functions of the molecular graph. Examples include eigenvalues of the adjacency matrix and descriptors based on the molecular influence matrix[20] etc. However, there are other descriptors that consider features such as paths and vertex degrees. Topological descriptors are able to provide a more detailed description of molecular shape features such as branching and crowdedness.

As an example consider the calculation of some connectivity indices. These were first described by Randic[21] and later extended by Kier and Hall[15] with the assumption that

> there resides in the structural formula sufficient information so that an index, based upon non-empirical counts of atoms, can be calculated

Numerous connectivity indices have been defined. A well-studied example is the χ index. These indices consider the vertex degree of each atom in various subgraphs of the molecular graph. Thus the ${}^1\chi$ index is defined as

$$ {}^1\chi = \sum_{i \neq j} \frac{1}{\sqrt{\delta_i \delta_j}} \tag{3.1} $$

where δ_i and δ_j are the vertex degrees of two bonded atoms, i and j. However, the ${}^1\chi$ descriptor is simplistic since it only considers atoms connected by a single bond. To characterize a molecular structure on a larger scale, extended versions of the χ descriptor were defined. Thus the second order connectivity index, ${}^2\chi$, is computed by *dissecting*[16] a structure into 2-bond (i.e., 3 atoms) fragments. The value of the descriptor is then calculated by

$$ {}^2\chi = \sum_{i \neq j \neq k} \frac{1}{\sqrt{\delta_i \delta_j \delta_k}} \tag{3.2} $$

where δ_i, δ_j and δ_k are the vertex degrees of the atoms in a given fragment. Higher order χ indices can be calculated in a similar manner. An important feature of these descriptors is that they restrict themselves to linear paths. However, structures exhibit branching and cyclic paths and the connectivity indices were extended to take these features into account, resulting in the ${}^m\chi_f$ descriptors where m denotes the number of edges in a fragment and f denotes the type of fragment that may be p (path), c (cluster), pc (path–cluster) or ch (chain). The structures of these fragments are summarized in Fig. 3.1.

It should be noted that the original definition only considered saturated carbon atoms. To take into account unsaturation and heteroatoms the δ value for an atom was modified such that

$$ \delta^v = Z^v - h \tag{3.3} $$

where Z^v is the number of valence electrons and h is the number of hydrogens. Eq. 3.3 was also extended to take into account core electrons for higher row elements. The use of the δ^v values lead to the calculation of the valence corrected χ indices denoted by ${}^m\chi_f^v$, where m and f have been previously defined.

Fig. 3.2 shows an example of how a molecular structure is decomposed into a variety of fragments. The original molecule (1-methyl 3-ethyl benzene) is decomposed into a 6th order chain (top), two 3rd order clusters (right) and three 4th order path–clusters (bottom). The numbers in the central structure correspond to the vertex degree for each atom. Thus the value of ${}^6\chi_{ch}$ would be obtained by

$$\begin{aligned} {}^6\chi_{ch} &= 1 \cdot \frac{1}{\sqrt{2 \times 2 \times 2 \times 2 \times 2 \times 2}} \\ &= 0.125 \end{aligned}$$

Similarly the fourth order path–cluster χ index, ${}^4\chi_{pc}$ would be calculated as

$$\begin{aligned} {}^4\chi_{pc} &= 3 \cdot \frac{1}{\sqrt{(1 \times 3 \times 2 \times 2 \times 2)}} \\ &= 0.612 \end{aligned}$$

Finally the value of the third order cluster χ index, ${}^3\chi_c$ would be calculated as

$$\begin{aligned} {}^3\chi_c &= 2 \cdot \frac{1}{\sqrt{(1 \times 3 \times 2 \times 2)}} \\ &= 0.576 \end{aligned}$$

Another type of topological descriptors are the BCUT's developed by Pearlman et al.[22] These descriptors are based on the Burden matrix[23] which is an adjacency matrix in which the non-diagonal elements are weighted based on the nature of the connectivity of the atoms involved. Thus for a molecule with n atoms and an $n \times n$ adjacency matrix, A, the Burden matrix, B is defined as

$$B_{ij} = \begin{cases} \pi \times 0.1 & \text{if } i \neq j \text{ and } A_{ij} = 1 \\ 0.001 & \text{if } i \neq j \text{ and } A_{ij} = 0 \\ Z & \text{if } i = j \end{cases}$$

where π represents the conventional bond order and Z represents the atomic number. Furthermore, all off-diagonal elements of B are augmented by 0.01.

The fundamental modification made by Pearlman was to place atomic properties along the diagonal of the Burden matrix. This leads to a variety of weighted Burden matrices where the weights include atomic weight, polarizability, electronegativity and hydrogen bonding ability. The actual descriptors are obtained by performing an eigenvalue decomposition of the Burden matrix and taking the lowest and highest eigenvalues. It has been shown that the extreme eigenvalues of the Burden matrix encode global information[24] regarding the molecule. Thus by combining atomic properties with the Burden matrix, the resultant eigenvalues encode global structure-property characteristics of a molecule, leading to BCUT descriptors being termed *holistic*. The holistic nature of these descriptors have led to their frequent use in studies of chemical diversity,[25–27] library design[28,29] and hit selection in high throughput screens.[30–32]

Topological descriptors have been widely used and have been shown to be very useful in building predictive models. Since they only require connectivity information for a molecule, the process of drawing and optimization of 3-D structures can be avoided. This results in the rapid calculation of this class of descriptors. The downside to topological descriptors is the lack of physical interpretability. Many of these types of descriptors are quite abstract in nature and though a number of reports have described correlations between certain descriptors and physical properties,[33–35] these are not easily generalized.

3.2.3 Electronic Descriptors

Electronic descriptors consider various features of the molecules' electronic environment. These include the HOMO and LUMO energies, electronegativity and various atom-centered partial charge descriptors. The ADAPT system is able to calculate atomic charges using empirical data to fit the dipole moments of molecules[36,37] or by using pK_a values.[38] These approaches are attractive since they do not require any optimization to be carried out and only consider molecular connectivity. The downside of these methods is that they are based on a predefined set of parameters and thus will not necessarily be accurate for a number of molecules. An alternative approach is to use an *ab initio* or semi-empirical technique to calculate charges. Owing to the time intensive nature of the former method, the semi-empirical approach is preferred and ADAPT is able to import partial charges calculated using the AM1[5] Hamiltonian with the MOPAC package. Though ADAPT focuses on charge based descriptors derived semi-empirically, there are a number of studies describing the development and application of *ab inito* quantum mechanical molecular descriptors[39–42] that calculate properties such as electron density, Fukui functions and so on. Owing to the computationally intensive nature of the calculation of these descriptors, a method has been developed that works with *atomic fragment values*, termed Transferable Atom Equivalent's (TAE)[43] which allow for the calculation of quantum mechanical descriptors for whole molecules using atomic fragment values. Together with a wavelet based encoding and a hybrid shape-property descriptor, TAE's have been used to build predictive QSAR models of high quality.[44]

3.2.4 Hybrid Descriptors

Hybrid descriptors are generally combinations of electronic or topological descriptors and geometric descriptors and in general characterize the distribution of a molecular feature over the whole molecule. Examples include the charged partial surface area (CPSA),[45] hydrophobic surface area (HPSA)[46] and hydrogen bonding[47,48] descriptors. The important characteristic is that they provide localized information regarding molecular features. Thus, in the case of the HPSA descriptors, one is able to obtain specific values of the hydrophobicity for different regions of the molecule as well as a global value for the whole molecule. For example, consider Fig. 3.3. In the upper figure, the atomwise hydrophobicity values are displayed. These hydrophobicity values are then color coded and mapped to the molecular surface to provide a visual representation of the information in the lower figure. The atomwise hydrophobicity values can be combined with surface area information for the individual atoms, as shown in Table 3.2, to obtain a wide variety of descriptors (25 in the ADAPT implementation). Examples include the atomic constant weighted hydrophobic and hydrophilic surface areas, total hydrophobic constant weighted hydrophobic surface area and the relative hydrophobicity. The functional forms of these descriptors are given below.

$$\begin{aligned}
\text{PPHS-2} &= \sum (+SA_i)(+\log P_i) \\
\text{PNHS-2} &= \sum (-SA_i)(-\log P_i) \\
\text{THWS} &= \sum (\log P_i)(SA_i) \\
\text{RPH-1} &= \frac{\text{Most hydrophobic atom constant}}{\sum \log P_i}
\end{aligned}$$

where SA_i is the surface area for the i^{th} atom and $\log P_i$ is the hydrophobic constant for the i^{th} atom and the $+$ and $-$ symbols indicate a hydrophobic or hydrophilic atom respectively.

The CPSA descriptors are similar in concept to the HPSA descriptors. In this case, the surface area values are combined with partial charges leading to 25 descriptors. In addition, a number of CPSA descriptors specific to certain atoms (such as N and O) are also calculated. These descriptors are similar in concept to the Polar Surface Area (PSA) descriptors[49, 50] which have been shown to be very useful in studies of intestinal absorption[51] and blood brain barrier crossing.[52] The development of the Topological Polar Surface Area (TPSA) method by Ertl et al.[53] allows the rapid evaluation of polar surface areas using only connectivity information (SMILES strings) and a library of fragment contributions.

By combining molecular surfaces with atomic properties, these descriptors are useful both in 2-D as well as 3-D QSAR methods. In addition, surface-property descriptor types usually have simple physical interpretations and have been shown to be quite information rich.[54, 55]

3.3 QSAR Set Generation

An important step in the modeling process is the creation of QSAR sets. Given a dataset of molecules, three mutually exclusive sets are created. The first, termed the training set, is used during the model building process. The learning algorithm used to build the model uses this set to characterize the dataset based on features present in the training set. The next set is the cross-validation set and is used in the case of CNN models. This set is used periodically during the training of the CNN and allows for the monitoring of the error rate during training. In the case of linear models the training set and cross-validation set are combined together. Finally the prediction set is a subset of the dataset that is not used at all during model building. Its purpose is to validate the final model and ascertain its predictive ability. These three sets are collectively termed QSAR sets.

The most common technique to generate these sets is random selection. The technique used in ADAPT is termed activity-weighted binning. In this procedure the dataset is binned based on activity values and then molecules are selected based on a probability, weighted by the bin populations.

An important point to note is that the learning algorithms are attempting to capture features of the dataset from a smaller subset of the overall dataset. When a model has been built it is tested on a another subset of the dataset. Clearly if the features that are present in the training set are not sufficiently represented in the prediction set, the models' predictive ability will be poor. Thus we must consider the idea of *representative* QSAR sets. In general, the QSAR sets should be created such that the various features present in the dataset should be proportionally represented in each individual QSAR set. One approach to this problem is to classify the dataset based on features described by a set of global molecular descriptors. The aim of such an approach is that these descriptors should be able to represent the main features of the dataset. The QSAR sets are then created such that molecules from the classes are represented in the same proportion that was found in the overall dataset. This approach is discussed in more detail in Chapter 4 where a Kohonen self organizing map is used to classify the dataset and subsequently create the QSAR sets. Alternative methods include the use of statistical molecular,[56] D-Optimal[57] or Kennard-Stone[58] design methods.

3.4 Feature Selection

Though only four types of descriptors has been mentioned above, these classes account for the nearly 300 descriptors calculated by ADAPT. Other programs such as DRAGON are able to evaluate nearly 1200 descriptors covering a wide variety of descriptor classes. It is apparent that in such a large descriptor pool a number of descriptors will be highly correlated with other descriptors or else may have the same value for all the molecules (such as number of aromatic rings, when the dataset has no aromatic rings) and will thus contain no relevant information. Thus before descriptors can be used for model building, the original descriptor pool must be reduced in size by selecting only feature rich and relevant descriptors. This selection step is termed objective feature selection. Once a reduced pool of descriptors has been created, suitable subsets of the descriptors must be selected to build QSAR models with. This step is termed subjective feature selection.

3.4.1 Objective Feature Selection

The original descriptor pool obtained from the evaluation of all available descriptors is reduced in size by two main methods. First, an identical test is carried out. This procedure removes descriptors that have a constant value for a user specified percentage of the dataset. In general the percentage ranges from 80% to 90%. The next step is to calculate the correlation coefficient between all the pairs of descriptors. If a pair of descriptors exhibit a R^2 value greater than or equal to a user specified cutoff, one member of the descriptor pair is discarded. Which pair is discarded is in general random; however, if one member of the pair is a topological descriptor, it is kept in preference to the other member. The reason for this behavior is that topological descriptors generally provide a global description of a molecule. Though the same is true of geometric descriptors, the larger number of topological descriptors available warrant their preferred inclusion in the reduced descriptor pool. Another technique that is used to create a reduced pool of descriptors is *vector space descriptor analysis*, which is based on the Gram-Schmidt orthogonalization procedure.[59] This technique considers descriptors as vectors and attempts to create a descriptor pool as a spanning linear vector space. Essentially, it starts by placing the descriptor that is most correlated to the dependent variable in the reduced pool. The next step is to find the descriptor from the original pool that is most orthogonal to the current descriptor. This step is repeated, each time selecting the descriptor from the original pool that is most orthogonal to the subspace spanned by the previously selected descriptors. The procedure is repeated until the number of descriptors in the reduced pool reaches a user-defined limit.

By varying the cutoffs for the identical and correlation tests and the size limit for the vector space technique, the size of the reduced descriptor pool can be varied by the user. In general a rule of thumb is used to decide on the size of the final reduced pool and is given by

$$\frac{n_{\mathrm{mol}}}{n_{\mathrm{reduced}}} = .6 \tag{3.4}$$

where n_{mol} is the number of molecules in the dataset and n_{reduced} is the number of descriptors in the reduced pool. This rule is derived from work carried out by Topliss et al.,[60] which quantitatively measured the relationship among the number of variables, the number of observations and the probability of chance correlations in linear regression models based on simulated data. The value of 0.6 represents a tradeoff between the numbers of variables and observations to minimize the probability of chance correlations.

3.4.2 Subjective Feature Selection

This stage of feature selection refers to methods by which descriptor subsets are selected from the reduced descriptor pool for model building purposes. The problem is combinatorial in nature; for reduced pools of moderate size a brute force approach to subset selection is unwieldy and for larger pools, computationally unfeasible. As a result, for reduced pools containing more than 20 descriptors stochastic search methods are preferred. Such methods include genetic algorithms[61,62] (GA), simulated annealing[63]

(SA), particle swarms[64] and ant colony algorithms.[65] ADAPT implements GA and SA methods for subjective feature selection.

The details of the GA and SA methods have been described in Chapter 2. The implementation of the GA in ADAPT involves the use of an objective function, which depends on the type of model being built. In the case of linear models the genetic algorithm is coupled with a linear regression routine. The fitness of a given descriptor subset is a function of the root mean square error (RMSE) of the model based on that subset. Another constraint that is sometimes applied is that models with values of the t-statistic less than 4.0 are rejected. However, it has been seen in practice that this sometimes leads to the rejection of models that have good predictive ability. Hence, this constraint is not strictly applied and models with lower values of the t-statistic are considered. In the case of descriptor subset selection for neural network models, the objective function is a 3-layer, fully-connected, feed-forward CNN as described in the previous chapter. The fitness for a given descriptor subset is defined using a cost function based on the RMS errors of the training and cross validation sets used in the CNN model. This cost function is defined as

$$\text{Cost} = \text{RMSE}_{TSET} + 0.5 \times |\text{RMSE}_{TSET} - \text{RMSE}_{CVSET}| \tag{3.5}$$

where RMSE_{TSET} and RMSE_{CVSET} are the RMSE values for the training and cross-validation sets, respectively. This cost function is designed to take into account model performance based on the training set as well as the extent of overfitting. As described in Chapter 2, care must be taken to prevent overfitting in a neural network model. This is controlled by the use of the cross-validation set. By considering the RMSE for the cross-validation set, the cost function penalizes models that cannot generalize as exhibited by having poor cross-validation performance. The constant factor of 0.5 is an empirically chosen value and has been observed to provide a balance between the RMSE values of the training and cross-validation sets.

In the case of the simulated annealing algorithm, the above discussion holds, except that the *energy* of a given configuration (i.e., descriptor subset) is now given by the RMSE (for linear models) or the value of the cost function (CNN models).

It should be noted that in the case of CNN models, the use of the genetic or simulated annealing algorithms results in models having optimal descriptor subsets for the specified architecture. To fully investigate the performance of a selected descriptor subset, a variety of CNN architectures must be considered. This is carried out by developing models with the same set of input descriptors but varying architectures (i.e., varying numbers of hidden layer neurons). The final model for a given descriptor subset is that which exhibits the lowest cost function.

3.5 Model Development

Once we have calculated the descriptors, reduced the original pool to a more manageable size and then selected a number of optimal descriptor subsets we can then proceed to build a set of models and choose the best one. The ADAPT methodology for model development involves three steps. First a set of linear models are developed using the top five to ten descriptor subsets selected by the GA or SA, coupled to the linear regression routine as the cost function. These models are termed Type I models. The best model is selected based on R^2 and RMSE value. In many cases, such as for biological properties, a simple linear relationship will not result in good predictive performance. Thus, the next step is to investigate whether the selected descriptor subset will show enhanced perfomance when used in a nonlinear relationship. Thus, we use the descriptor subset from the linear model and build a nonlinear CNN model. For the given descriptor subset (i.e., input neurons) a number of CNN models are developed by varying the number of hidden neurons, subject to the constraint specified by Eq. 2.5. Out of this set of models the final model is the one that exhibits the lowest cost function defined by Eq. 3.5. This model is termed a Type II model. The problem with this type of model is that it uses a descriptor subset that was selected by the GA (or SA), based on its performance in a linear model. That is, the descriptor subset was optimal for linear models but not necessarily for nonlinear models. As a result, the final step of model building consists of using the GA (or SA) coupled to the CNN routine to search for descriptor subsets that show good performance in CNN models. Once a number of descriptor subsets have been obtained, the final architecture is obtained as described above. Nonlinear models that are obtained by linking the feature selection routines to a nonlinear cost function are termed Type III models. The model development procedure described here is summarized graphically in Fig. 3.4. The result of this procedure is to create a set of linear and nonlinear models. In many cases, both types of models can be used in combination to investigate different aspects of the structure-property relationship being modeled and in other cases one type of model may be sufficient to understand the trends present in the dataset as well as provide good predictive ability for new observations.

3.6 Prediction, Validation and Interpretation

After a QSAR model has been developed the next step is to investigate its predictive ability. The simplest method is to test the model on a subset of the dataset that has not been used during the model development process (the prediction set). The statistics obtained from the results of the prediction set can give us some indication of the model's predictive ability. The most common statistics for linear models are R^2 and RMSE, though the former is not always a very reliable indicator of the goodness of fit as shown in Fig. 3.5. The figure plots the predicted versus observed values obtained from a linear regression model based on a simulated dataset. The dataset consisted of two well-seperated Gaussian clusters. Clearly, the relationship between the independent variables and the dependent variable is not linear. However, the R^2 value of 0.91 misleadingly indicates that the regression model fits the data well.

Another aspect closely related to predictive ability is *generalizability.* The main problem with the use of a single prediction set as a test of a model's predictive ability is that it is a limited indicator of the model's ability to handle new data. Generalizability is a more general term than predictive ability and essentially describes how the model behaves when faced with new data. The question of generalizability arises owing to the fact that a testing methodology based on a subset of the original dataset is inherently biased since the prediction set will, to some extent, share distribution characteristics of the training set. Obviously this may not always be true and is dependent on the manner in which the training and prediction sets are generated. But in general one can assume that new datasets will share the characteristics of the data to differing extents. Clearly a new dataset that differs greatly from the training data (say a dataset of linear molecules versus a dataset of cyclic molecules) will not give rise to good predictions from the model. On the other hand a new dataset containing molecules that are similar to the training data can be expected to lead to good predictions.

How can we measure generalizability? The answer to this question is not clear cut. One possible indicator of model generalizability is the relative performance of the model on cross-validation and prediction sets. This possibility is discussed in more detail in Chapter 4. An important point to note regarding this approach is that this requires the use of a cross-validation set and consequently cannot be applied directly to linear models built using multiple linear regression. An alternative approach to the question of generalizability, alluded to above, is to try and quantify how well a new dataset will be predicted by a model. Essentially, this method tries to link some aspect of model quality to the structures of the molecules being considered. One approach is to link model quality to some similarity measure between the training dataset and a new dataset. Yet another possibility is to predict the performance of a model on a new dataset directly, using information from the model and the new structures. These approaches are discussed in more detail in Chapter 5.

Validation of a QSAR model is very similar in nature to the ideas discussed above. However, whereas the above discussion focuses on validation of a final QSAR model, validation also plays an important role during model development and is generally termed cross-validation. More specifically, algorithms such as neural networks and random forests all benefit from a validation mechanism during model development. In the case of a neural network, cross-validation is required to prevent over-training as described in Section 2.2.1. Similarly, the random forest algorithm uses a built-in cross-validation scheme to provide an internal measure of accuracy.

The ADAPT CNN methodology uses two forms of validation. One option is to use a fixed cross-validation set through all validation iterations. The second option is to use a leave n% out validation scheme. The latter method works by randomly selecting n% of the training set at each validation iteration and evaluating a cross-validation RMSE. A wrapper is also available which carries out a *round robin* leave n% out validation scheme (though this is probably more correctly termed as an ensemble method). In contrast to the above method it generates multiple training, cross-validation and prediction sets such that each member of the dataset is present in one of the prediction sets (and correspondingly one of the cross-validation sets). Though this is more rigorous than than a neural network algorithm with a fixed cross-validation set it is probably not as useful as the leave n% out method using randomly selected cross-validation sets. The reasons are twofold. First, the procedure whereby each member of the dataset is predicted once is extremely time consuming. Second, this procedure results in an ensemble of neural network models (for each training, cross-validation and prediction set combination) rather than a single model. One possible justification for ignoring the latter drawback is that neural network models are in general not considered interpretable and are usually developed for their predictive ability. Thus the fact that an ensemble of models is generated rather than a single model may be justified to some extent if the predictive ability of the ensemble is significantly better than the single model. However, as noted by Agrafiotis et al.,[66] the “benefits of aggregation methods are clear but not overwhelming.”

Validation, in the sense of neural networks and random forests, is not directly applicable to the case of linear model’s developed using multiple linear regression. However, one method that can be used to gain an idea of a linear model’s predictive ability is to use a leave-one-out (LOO) procedure resulting in a prediction for each member of the dataset. This method results in a cross-validated R^2, usually denoted by Q^2. However, the utility of this statistic is debatable and numerous discussions are available in the literature.[67–71]

An important component of the validation process is testing for chance correlations. That is, we would like to know whether the results generated by the model were due to chance correlations rather than the model actually capturing a specific structure activity relationship (SAR). This is important in the context of the ADAPT methodology as the algorithms used during subjective feature selection are stochastic in nature. Thus it is possible that the results of a model developed on the basis of a descriptor subset selected by the GA or SA are simply due to luck rather than an any real relationship between the dependent variable and the independent variables. The simplest strategy to test for chance correlations is to scramble the dependent variable and estimate R^2 and RMSE values for the model using the scrambled dependent variable. Since the fundamental assumption of QSAR modeling is that descriptor values correlate with the observed activity (or property) one would expect that the R^2 for the scrambled dependent variable would decrease and that the RMSE would increase. Graphically, a plot of the observed versus predicted property should appear random as illustrated in Fig. 3.6. If the results of the scrambled runs are similar to those produced by a model using the true dependent variable then one must conclude that the model has not captured a real structure activity relationship. Topliss et al.[60] discuss the role of chance correlations in the context of linear regression and their simulations provide a guide to the probability of observing a given value of R^2 (for the case of random variables). The simulation only considered a small set of possible variable combinations and thus is not an exhaustive study. However it does indicate the importance of checking for chance correlations. The method of scrambling the dependent variable can be applied to both linear and nonlinear models. In either case this technique tests the resultant model for chance correlations.

Another possibility is to test the feature selection algorithms themselves for chance correlations. That is, are the best descriptor subsets arising due to chance or are they really minima in the descriptor space searched by the GA or SA? A simple way to investigate this type of chance correlations is to evaluate the statistics of models built from randomly selected descriptors. Similar results as described above would be expected. In this case the difference should not be as large since the descriptors will still be correlated to the dependent variable, but owing to random selection the descriptor combinations may not be optimal and hence should result in poorer statistics compared to a model built with an optimal subset of descriptors.

At this point we have in hand a validated model with (it is hoped) good predictive ability. The important feature of the model is that it should have incorporated one or more structure activity relationships. The final task of a QSAR modeling methodology is to interpret the model to describe these relationships. The ADAPT methodology leads to both linear and non-linear models and currently both types of models can be interpreted. The interpretation of linear models utilizes the PLS technique described by Stanton.[72] Its ability to dissect the effects of individual descriptors on the dataset allows a very detailed description of any structure activity relationship captured by the model. A brief description of the PLS technique is provided below.

The first requirement of this technique is to have a statistically valid linear regression model - generally characterized by high absolute values of individual t-statistics and a high value of the overall F-statistic. The next step is to build a PLS model using the selected descriptors. An important observation at this point is that the PLS algorithm employed in this work used a leave-one-out cross-validation scheme to determine the optimal number of PLS components. If the optimal number indicated by cross-validation does not equal the number of descriptors, the initial linear model was overfit and thus cannot be usefully analyzed by the PLS technique.[72] Given a validated model we extract the X and Y scores and the X weights from the PLS analysis. The X weights give an $m \times n$ matrix, where m is the number of descriptors and n is the number of PLS components, which are simply linear combinations of the descriptors used in the original original linear models. Essentially each column can be interpreted as the contributions of individual descriptors to a given component. The X and Y scores will be also be matrices with the PLS components in the columns and the observations in the rows. The Y score vector for a given component is analogous to a predicted value made by the original linear model, except that now it models the transformed variable denoted by the X score vector. For each component we create scoreplots by plotting the X score vector against the Y score vector. The next stage involves a simultaneous analysis of the components and their corresponding scoreplots. Ideally we would see that there are one or two descriptors in each component that have high weight values - indicating that they are the main contributors to the component. We start with the first PLS component and obtain the most weighted descriptor. We then consider the score plot for that component. Compounds in the upper right and lower left are properly predicted whereas compounds lying in the other quadrants are either over-predicted (upper left) or under-predicted (lower right). Compounds that are correctly predicted as active will tend to lie in the upper right quadrant and those that are correctly predicted as inactive will occupy the lower left quadrant of the scoreplot. One can thus conclude that compounds with high values of the most weighted descriptor (assuming the weight is positive) will be more active than compounds with low values. This argument is reversed if the weight for the descriptor is negative. The under- or over-predicted compounds are not explained by the current component. Thus we must consider the next component and its most weighted descriptor. One would expect that compounds that were poorly predicted by the first component will be well predicted by the second one and the most weighted descriptor for this component will be able to account for the good predictions. Once again, for the poorly predicted cases, we move to the next component and proceed as before.

At the end of this procedure the role of the individual descriptors in determining activity (or lack of it) will have been extracted from the model. In the words of Stanton,[73] “it’s like reading a book”. This technique has been used in the interpretation of biological activity of artemisinin analogous[74] and the inhibitory activity of PDGFR inhibitors.[55]

In the case of a neural network model two forms of interpretation can be generated. First, a measure of the importance of the input descriptors can be generated using a technique analogous to the measure of variable importance in random forests.[75,76] In addition, we can also provide a more detailed interpretation of a CNN model based on a method inspired by the PLS technique described above. The development of the CNN interpretation methodologies and examples of applications are described in Chapters 8 and 9.

3.7 Conclusions

The development of QSAR models proceeds in a stepwise fashion as described in this chapter. The first step is the entry of the molecular structures and optimizations for geometry and electronic properties. Next, molecular descriptors are calculated for the dataset and objective feature selection is carried out to reduce the number of descriptors to a manageable pool. The next step is to select subsets of descriptors to build models. As has been shown, descriptor selection and model building are interlinked, using stochastic algorithms to search for descriptor subsets that lead to low cost (in terms of RMS error for linear and cost function for nonlinear) models.

The model building process generally proceeds in three stages. In the first stage a set of linear models are built. In the second stage, the descriptor subsets used in the best linear models are then used to build neural network models, the assumption being, that, if a nonlinear structure-activity relationship is present, the CNN should be able to better capture it. In the third stage, the GA or SA feature selection method is coupled with the CNN routine to search for descriptor subsets that perform optimally in a nonlinear model. In both the second and first phases, the final architecture of the CNN model, for a selected descriptor subset, is decided upon by rigorously investigating all possible architectures subject to the constraint on the number of adjustable parameters.

Finally, after a number of models have been generated, they are validated and then investigated for predictive and interpretive ability. The former is usually good for the selected models. The resultant models can then be interpreted. Depending on the type of model different degrees of interpretation are possible. Linear regression and CNN models can be interpreted in a detailed manner. In addition, broad measures of descriptor importance can also be obtained for CNN models and ensemble models (such as random forest models) though such interpretations are necessarily not as informative.

The following chapters discuss applications of the QSAR methodology described here as well as investigations of specific steps in the QSAR methodology.

Table 3.1: A list of the descriptors and their associated class available in ADAPT

Type	Name	Function	Reference
Topological	DKAPPA	κ shape indices	77–79
	DMALP	All self avoiding paths of length upto the longest path in the structure	80, 81
	DMCHI	χ molecular connectivity indices	15–17
	DMCON	Molecular connectivity indices, similar to DMCHI but corrects for heteroatoms in rings and aromatic rings	82, 83
	DMFRAG	Counts for a variety of substructures	
	DMWP	Weighted paths based on Randic's molecular ID	21
	DEDGE	Molecular distance edge descriptor, λ	18
	CTYPES	Hybridization of carbon atoms based on connectivity only	
	DESTAT	Electrotopological state	84, 85
	DPEND	Superpendentic index	86
Geometric	DSYM	Structural symmetry index, equal to ratio of the number of unique atoms to the total number of atoms in a hydrogen suppressed structure	
	ECCEN	Eccentric connectivity index	19
	DMOMI	Moments of inertia along X, Y and Z axes	11
	SAVOL	Molecular surface area and volume	12
	SHADOW	Shadow areas obtained by projecting a 3-D structure onto the XY, XZ or YZ planes	13, 14
	DGRAV	Gravitational index	87
	LOVERB	Molecular length to breadth ratio	
Electronic	CHARGE	Dipole moment, charges on most negative and positive atoms and the sum of absolute values of all charges	
	HLEH	HOMO & LUMO energies and electronegativity and hardness	
	MRFRAC	Molecular refraction	88
	MPOLR	Molecular polarizability	89
Hybrid	CPSA	Charged partial surface areas	45

Table 3.1: (continued)

Type	Name	Function	Reference
	DATOM	CPSA descriptors for specific groups and atoms (carbonyl, O, N, S, and halogens)	
	HBSA	Hydrophobic surface areas	54, 90
	HBMIX	Intermolecular hydrogen bonding ability	47, 48
	HBPURE	Intramolecular hydrogen bonding ability	47, 48

Table 3.2. The hydrophobicity and solvent accessible surface area values calculated for glycine. These values are combined to generate the 25 HPSA[54] descriptors.

Serial No.	Atom Label	Hydrophobicity	Surface Area (Å^2)
1	C	-0.20	2.58
2	C	-0.28	7.36
3	O	-0.15	48.00
4	O	-0.29	26.00
5	N	-1.02	19.04
6	H	0.12	25.01
7	H	0.12	25.33
8	H	0.30	30.81
9	H	0.21	29.81
10	H	0.21	21.98

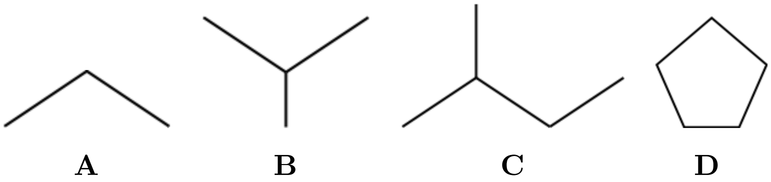

Fig. 3.1. The four types of fragments used to calculate χ descriptors. **A** – 2nd order path. **B** – 3rd order cluster. **C** – 4th order path cluster. **D** – 5th order chain. The order refers to the number of edges in each fragment.

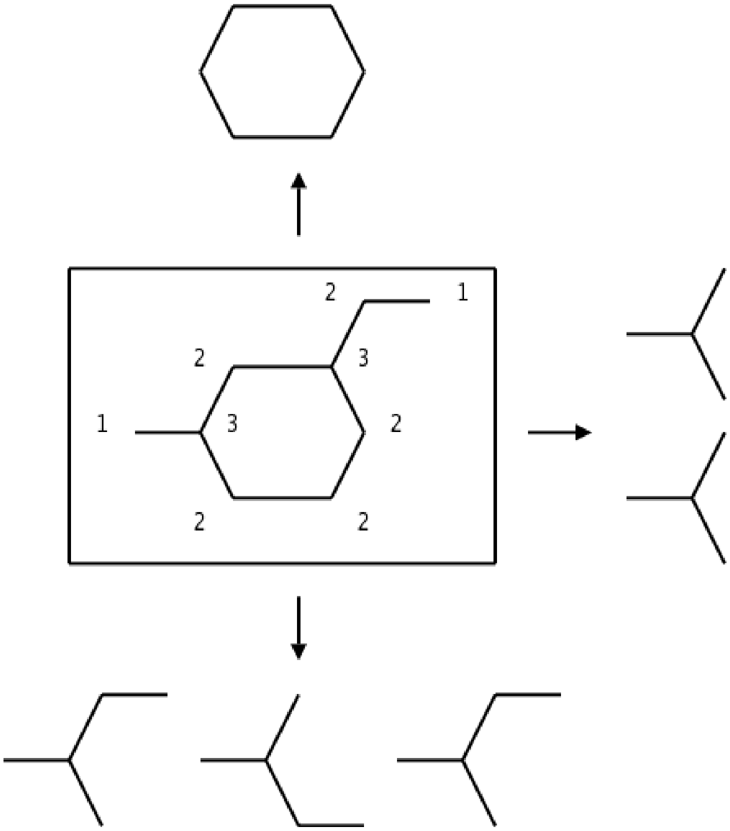

Fig. 3.2. A diagram illustrating the decomposition of 1-methyl 3-ethyl benzene into fragments for subsequent use in the calculation of χ descriptors. The annotations of the central structure correspond to the vertex degree of each atom.

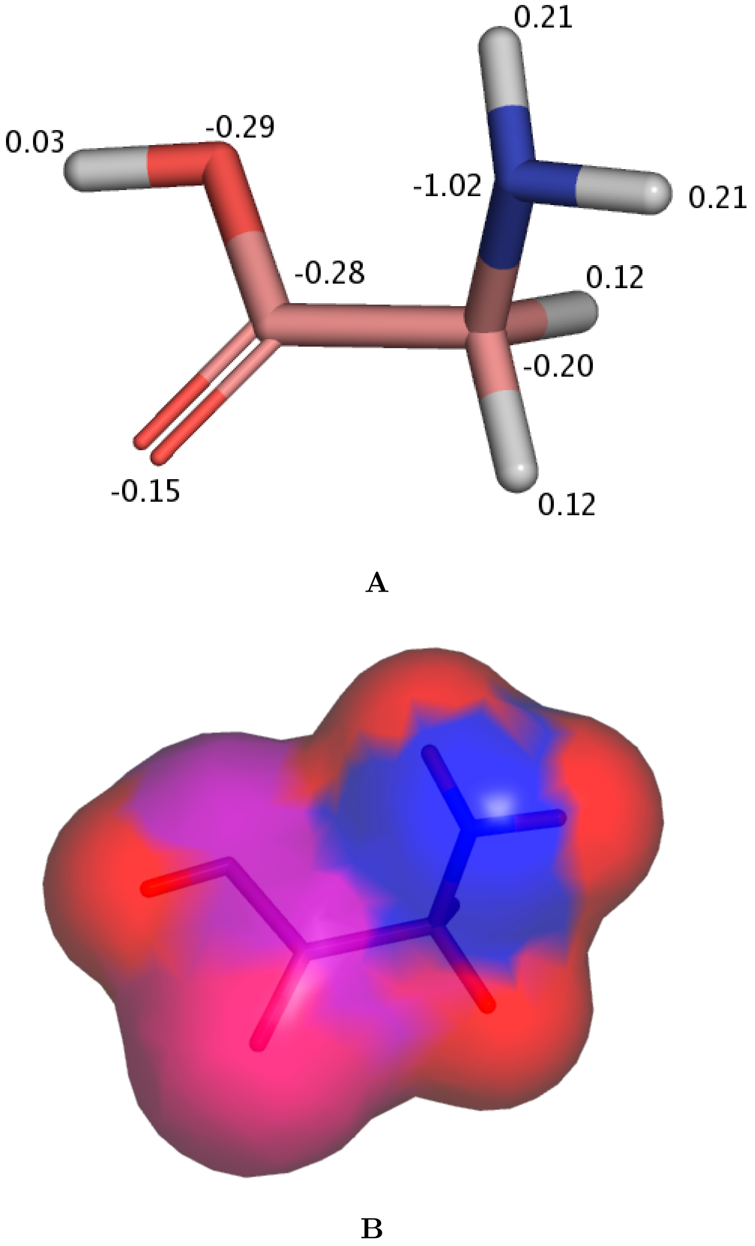

Fig. 3.3. Graphical representations of hydrophobicity values for the glycine molecule. **A** shows the numerical hydrophobicity values and **B** displays the solvent accessible surface area color coded by the hydrophobicity values. Blue regions indicate the most hydrophilic groups and red corresponds to the most hydrophobic groups.

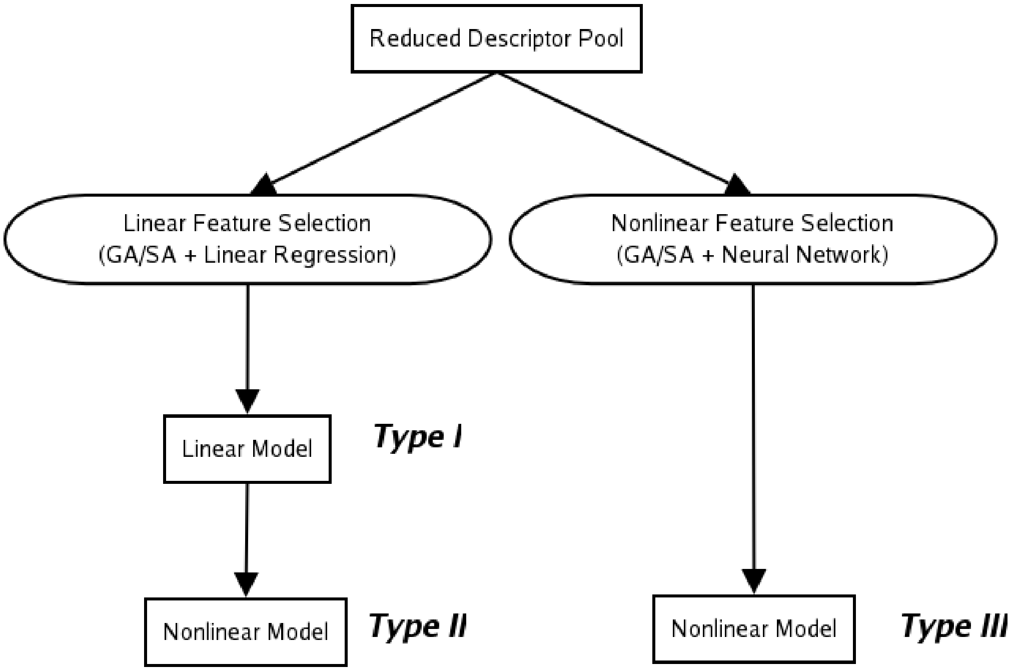
Reduced Descriptor Pool
Linear Feature Selection (GA/SA + Linear Regression)
Nonlinear Feature Selection (GA/SA + Neural Network)
Linear Model
Type I
Nonlinear Model
Type II
Nonlinear Model
Type III

Fig. 3.4. The sequence of steps involved in model building using the ADAPT methodology. Here GA and SA refer to the genetic algorithm and simulated annealing feature selection methods respectively.

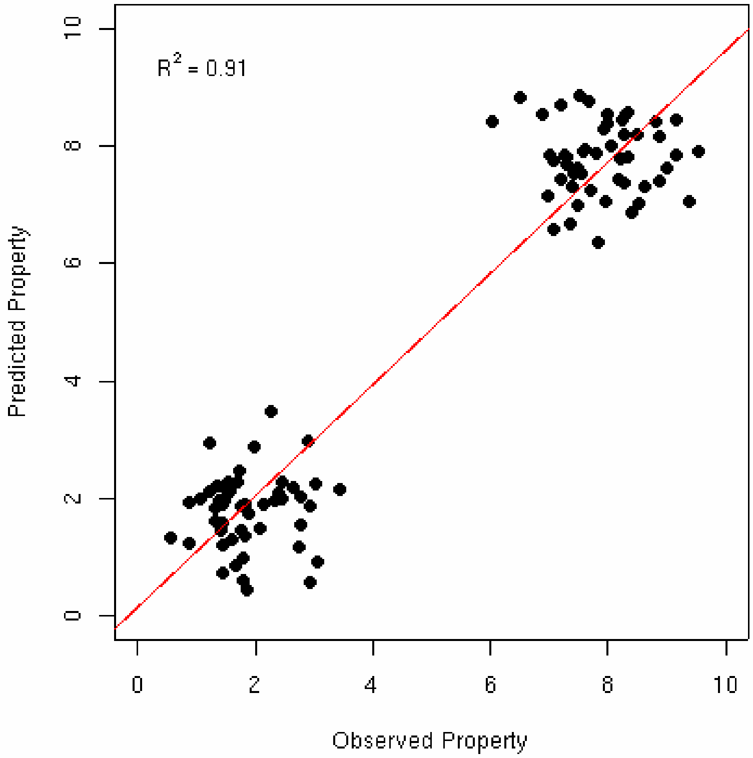

Fig. 3.5. A plot generated from a linear regression model, using simulated data, with a high value of R^2, but clearly unable to explain the variation in the dataset. The red line represents the fitted regression line.

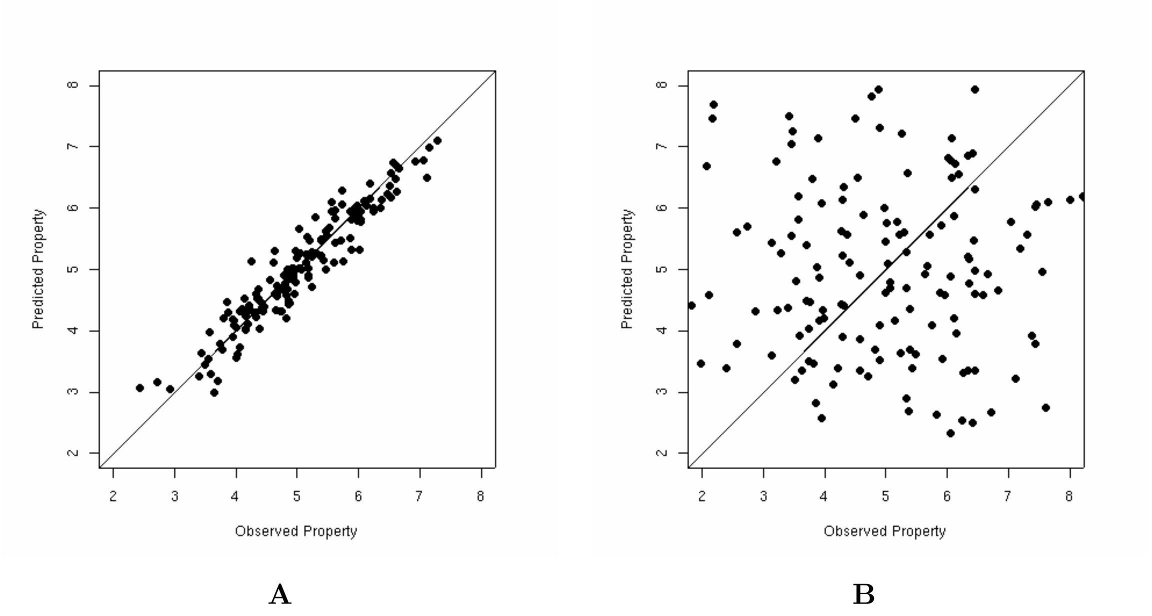

Fig. 3.6. Plots generated using simulated data, illustrating the results of testing chance correlations in a linear model by scrambling the dependent variable. Plot **A** represent the original linear model. Plot **B** represents the linear model rebuilt after scrambling the independent variable.

References

[1] Jurs, P.; Chou, J.; Yuan, M. Computer Assisted Drug Design. In ; American Chemical Society: Washington D.C., 1979; Chapter Studies of Chemical Structure Biological Activity Relations Using Pattern Recognition.

[2] Stuper, A.; Brugger, W.; Jurs, P. *Computer Assisted Studies of Chemical Structure and Biological Function;* Wiley: New York, 1979.

[3] Stewart, J. Optimization of Parameters for Semi-Empirical Methods I – Method. *J. Comp. Chem.* **1989,** *10,* 209–220.

[4] Stewart, J. Optimization of Parameters for Semi-Empirical Methods II – Applications. *J. Comp. Chem.* **1989,** *10,* 221–64.

[5] Dewar, M.; Zoebisch, E.; Healy, E.; Stewart, J. AM1 – A New General Purpose Quantum Mechanical Molecular Model. *J. Am. Chem. Soc.* **1985,** *107,* 5348–5348.

[6] Davis, L.; Burggraf, L.; Storch, D. Hydration of Small Anions - Calculations by the AM1 Semi-Empirical Method. *J. Comp. Chem.* **1991,** *12,* 350–358.

[7] Dewar, M.; McKee, M.; Rzepa, H. MNDO Parameters for Third Period Elements. *J. Am. Chem. Soc.* **1978,** *100,* 3607.

[8] Todeschini, R.; Consonni, V.; Pavan, M. “DRAGON”, .

[9] Wegner, J. “JOELib”, `http://joelib.sf.net`, 2005.

[10] Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors;* Wiley-VCH: Berlin, 2002.

[11] Goldstein, H. *Classical Mechanics;* Addison Wesley: Reading, MA, 1950.

[12] Pearlman, R. Physical Chemical Properties of Drugs. In ; Marcel Drekker, Inc.: New York, 1980; Chapter Molecular Surface Area and Volumes and their Use in Structure-Activity Relationships.

[13] Stouch, T.; Jurs, P. A Simple Method for the Representation, Quantification, and Comparison of the Volumes and Shapes of Chemical Compounds. *J. Chem. Inf. Comput. Sci.* **1986,** *26,* 4–12.

[14] Rohrbaugh, R.; Jurs, P. Molecular Shape and Prediction of High Performace Liquid Chromatographic Retention Indices of Polycyclic Aromatic Hydrocarbons. *Anal. Chem.* **1987,** *59,* 1048–1054.

[15] Kier, L.; Hall, L.; Murray, W. Molecular Connectivity I: Relationship to local anasthesia. *J. Pharm. Sci.* **1975,** *64,* 1971–1974.

[16] Kier, L.; Hall, L. *Molecular Connectivity in Structure Activity Analysis;* John Wiley & Sons: Hertfordshire, England, 1986.

[17] Kier, L.; Hall, L. Molecular Connectivity VII: Specific Treatment to Heteroatoms. *J. Pharm. Sci.* **1976,** *65,* 1806–1809.

[18] Liu, S.; Cao, C.; Li, Z. Approach to Estimation and Prediction for Normal Boiling Point (NBP) of Alkanes Based on a Novel Molecular Distance Edge (MDE) Vector, λ. *J. Chem. Inf. Comput. Sci.* **1998,** *38,* 387–394.

[19] Sharma, V.; Goswami, A.; Madan, A. Eccentric Connectivity Index: A Novel Highly Discriminating Topological Descriptor for Structure-Property and Structure-Activity Studies. *J. Chem. Inf. Comput. Sci.* **1998,** *37,* 273–282.

[20] Consonni, V.; Todeschini, R.; Pavan, M. Structure-Response Correlations and Similarity/Diversity Analysis By GETAWAY Descriptors. Part 1. Theory of the Novel 3D Molecular Descriptors. *J. Chem. Inf. Comput. Sci.* **2002,** *42,* 682–692.

[21] Randic, M. On Molecular Idenitification Numbers. *J. Chem. Inf. Comput. Sci.* **1984,** *24,* 164–175.

[22] Pearlman, R.; Smith, K. 3D-QSAR in Drug Design. In ; Kubinyi, H. e. a., Ed.; Kluwer/Escom: Dordrecht, The Netherlands, 1998.

[23] Burden, F. Molecular Identification Number for Substructure Searches. *J. Chem. Inf. Comput. Sci.* **1989,** *29,* 225–227.

[24] Burden, F. A Chemically Intuitive Molecular Index Based on the Eigenvalues of a Modified Adjacency Matrix. *Quant. Struct. Act. Relat.* **1997,** *16,* 309–314.

[25] Pearlman, R.; Smith, K. Novel Software Tools for Chemical Diversity. *Persp. Drug Discov. Design* **1998,** *9,* 339–353.

[26] Stanton, D. Evaluation and Use of BCUT Descriptors in QSAR and QSPR Analysis.. *J. Chem. Inf. Comput. Sci.* **1999,** *39,* 11–20.

[27] Martin, Y. Diverse Viewpoints on Computational Aspects of Molecular Diversity. *J. Comb. Chem.* **2001,** *3,* 231–250.

[28] Schnur, D. Design and Diversity Analysis of Large Compound Libraries Using Cell-Based Methods. *J. Chem. Inf. Comput. Sci.* **1999,** *39,* 36–45.

[29] Young, S. S.; Wang, M.; Gu, F. Design of Diverse and Focused Combinatorial Libraries Using an Alternating Algorithm. *J. Chem. Inf. Comput. Sci.* **2003,** *43,* 1916–1921.

[30] Shanmugasundaram, V.; Maggiora, G.; Lajiness, M. Hit Directed Nearest-Neighbor Searching. *J. Med. Chem.* **2005,** *48,* 240–248.

[31] Stanton, D. T.; Morris, T. W.; Roychoudhury, S.; Parker, C. N. Application of Nearest-Neighbor and Cluster Analysis in Pharmaceutical Lead Discovery. *J. Chem. Inf. Comput. Sci.* **1999,** *39,* 21–27.

[32] Manallack, D. T.; Pitt, W. R.; Gancia, E.; Montana, J. G.; Livingstone, D. J.; Ford, M. G.; Whitley, D. C. Selecting Screening Candidates for Kinase and G Protein-Coupled Receptor Targets Using Neural Networks. *J. Chem. Inf. Comput. Sci.* **2002,** *42,* 1256–1262.

[33] Todeschini, R.; Cazar, R.; Collina, E. The Chemical Meaning of Topological Indices. *Chemom. Intell. Lab. Sys.* **1975,** *97,* 6609–6615.

[34] Randic, M.; Zupan, J. On Interpretation of Well Known Topological Indices. *J. Chem. Inf. Comput. Sci.* **2001,** *41,* 550–560.

[35] Randic, M.; Balaban, A.; Basak, S. On Structural Interpretation of Several Distance Related Topological Indices. *J. Chem. Inf. Comput. Sci.* **2001,** *41,* 593–601.

[36] Dixon, S.; Jurs, P. Atomic Charge Calculations for Quantitative Structure-Property Relationships. *J. Comp. Chem.* **1992,** *13,* 492–504.

[37] Abraham, R.; Griffiths, L.; Loftus, J. Approaches to Charge Calculations in Molecular Mechanics. *J. Comp. Chem.* **1982,** *3,* 407–416.

[38] Dixon, S. L. *Development of Computational Tools for Use in Quantitative Structure-Activity and Structure-Property Relationships,* PhD thesis, Department of Chemistry, Pennsylvania State University, 1994.

[39] Estrada, E.; Perdomo-Lopez, I.; Torres-Labandeira, J. J. Combination of 2D-, 3D-Connectivity and Quantum Chemical Descriptors in QSPR. Complexation of α- and β-Cyclodextrin with Benzene Derivatives. *J. Chem. Inf. Comput. Sci.* **2001,** *41,* 1561–1568.

[40] Netzeva, T. I.; Aptula, A. O.; Benfenati, E.; Cronin, M. T. D.; Gini, G.; Lessigiarska, I.; Maran, U.; Vracko, M.; Schuurmann, G. Description of the Electronic Structure of Organic Chemicals Using Semiempirical and Ab Initio Methods for Development of Toxicological QSARs. *J. Chem. Inf. Model.* **2005,** *45,* 105–114.

[41] Karelson, M.; Lobanov, V.; Katritzky, A. Quantum-Chemical Descriptors in QSAR/QSPR Studies. *Chem. Rev.* **1996,** *105,* 7512–7516.

[42] Thanikaivelan, P.; Subramnian, V.; Rao, J.; Nair, B. Application of Quantum Chemical Descriptors in Quantitative Structure Activity and Structure Property Relationships. *Chem. Phys. Lett.* **2000,** *323,* 59–70.

[43] Whitehead, C.; Sukumar, N.; Breneman, C. Transferable Atom Equivalent Multi-Centered Multipole Expansion Method. *J. Comp. Chem.* **2003,** *24,* 512–529.

[44] Breneman, C.; Sundling, C.; Sukumar, N.; Shen, L.; Katt, W. New Developments in PEST Shape/Property Hybrid Descriptors. *J. .Comp. Aided Mol. Des.* **2003,** *17,* 213–240.

[45] Stanton, D.; Jurs, P. Development and Use of Charged Partial Surface Area Structural Descriptors in Computer Assissted Quantitative Structure Property Relationship Studies. *Anal. Chem.* **1990,** *62,* 2323–2329.

[46] Stanton, D. T.; Mattioni, B. E.; Knittel, J. J.; Jurs, P. C. Development and Use of Hydrophobic Surface Area (HSA) Descriptors for Computer-Assisted Quantitative Structure-Activity and Structure-Property Relationship Studies. *J. Chem. Inf. Comput. Sci.* **2004,** *44,* 1010–1023.

[47] Pimentel, G.; McClellan, A. *The Hydrogen Bond;* Reinhold Pub. Corp.: New York, 1960.

[48] Vinogradov, S.; Linnell, R. *Hydrogen Bonding;* Van Nostrand Reinhold: New York, 1971.

[49] Palm, K.; Luthman, K.; Ungell, A.-L.; Strandlund, G.; Beigi, F.; Lundahl, P.; P., A. Evaluation of Dynamic Polar Molecular Surface Area as Predictor of Drug Absorption: Comparison with Other Computational and Experimental Predictors. *J. Med. Chem.* **1998,** *41,* 5382–5392.

[50] Stenberg, P.; Luthman, K.; Artursson, P. Prediction of Membrane Permeability to Peptides from Calculated Dynamic Molecular Surface Properties. *Pharm. Res.* **1999,** *16,* 972–978.

[51] Palm, K.; Stenberg, P.; Luthman, K.; Artursson, P. Polar Molecular Surface Properties Predict the Intestinal Absorption of Drugs in Humans. *Pharm. Res.* **1997,** *14,* 568–571.

[52] Clark, D. Rapid Calculation of Polar Molecular Surface Area and its Application to the Prediction of Transport Phenomena. 2. Prediction of Blood-Brain Barrier Penetration. *J. Med. Chem.* **1999,** *88,* 815–821.

[53] Ertl, P.; Rohde, B.; Selzer, P. Fast Calculation of Molecular Polar Surface Area as a Sum of Fragment-based Contributions and its Application to the Prediction of Drug Transport Properties. *J. Med. Chem.* **2000,** *43,* 3714–3717.

[54] Stanton, D.; Mattioni, B. E.; Knittel, J.; Jurs, P. Development and Use of Hydrophobic Surface Area (HSA) Descriptors for Computer Assisted Quantitative Structure-Activity and Structure-Property Relationships. *J. Chem. Inf. Comp. Sci.* **2004,** *44,* 1010–1023.

[55] Guha, R.; Jurs, P. C. The Development of Linear, Ensemble and Non-Linear Models for the Prediction And Interpretation of the Biological Activity of a Set of PDGFR Inhibitors. *J. Chem. Inf. Comput. Sci.* **2004,** *44,* 2179–2189.

[56] Andersson, P.; Lundstedt, T. Hierarchical Experimental Design Exemplified By QSAR Evaluation of a Chemical Library Directed Towards the Melanocortin–4 Receptor. *J. Med. Chem.* **2002,** *16,* 490–496.

[57] Wu, W.; Walczak, B.; Massart, D.; Heuerding, S.; Erni, F.; Last, I.; Prebble, K. Artificial Neural Networks In Classification of NIR Spectral Data: Design of the Training Set. *Chemom. Intell. Lab. Sys.* **1996,** *33,* 35–46.

[58] Kocjancic, R.; Zupan, J. Modelling of the River Flowrate: The Influence of the Training Set Selection.. *Chemom. Intell. Lab. Sys.* **2000,** *54*, 21–34.

[59] Arfken, G.; Weber, H. *Mathematical Methods for Physicists;* Harcourt/Academic Press: San Diego, CA, 2000.

[60] Topliss, J.; Edwards, R. Chance Factors in Studies of Quantitative Structure-Activity Relationships. *J. Med. Chem.* **1979,** *22,* 1238–1244.

[61] Cramer, N. A Representation for the Adaptive Generation of Simple Sequential Programs. In *Proc. of the International Conference on Genetic Algorithms and their Applications*; Lawrence Erlbaum Associates: Pittsburgh, PA, 1985.

[62] Goldberg, D. *Genetic Algorithms in Search, Optimization & Machine Learning;* Addison-Wesley: Reading, MA, 2000.

[63] Metropolis, N.; Rosenbluth, A.; Rosenbluth, M.; Teller, A.; Teller, E. Equation of State Calculations by Fast Computing Machines. *J. Chem. Phys.* **1953,** *21*, 1087–1092.

[64] Agrafiotis, D. K.; Cedeño, W. Feature Selection for Structure-Activity Correlation Using Binary Particle Swarms. *J. Med. Chem.* **2002,** *45*, 1098–1107.

[65] Izrailev, S.; Agrafiotis, D. A Novel Method for Building Regression Tree Models for QSAR Based on Artificial Ant Colony Systems. *J. Chem. Inf. Comput. Sci.* **2001,** *41,* 176–180.

[66] Agrafiotis, D.; Cedeño, W.; Lobanov, V. On the Use of Neural Network Ensembles in QSAR and QSPR. *J. Chem. Inf. Comput. Sci.* **2002,** *42,* 903–911.

[67] Golbraikh, A.; Tropsha, A. Beware of q^2. *J. Mol. Graph. Model.* **2002,** *20,* 269–276.

[68] Golbraikh, A.; Shen, M.; Xiao, Z.; Xiao, Y.; Lee, K.; Tropsha, A. Rational Selection of Training and Test Sets for the Development of Validated QSAR Models. *J. Comput. Aid. Mol. Des.* **2003,** *17*, 241–253.

[69] Kubinyi, H.; Hamprecht, F.; Mietzner, T. Three-Dimensional Quantitative Similarity-Activity Relationships (3D QSAR) from SEAL Similarity Matrices. *J. Med. Chem.* **1998,** *41,* 2553–2564.

[70] Novellino, E.; Fattorusso, C.; Greco, G. Use of Comparative Molecular Field Analysis and Cluster Analysis in Series Design. *Pharm. Acta. Helv.* **1995,** *70,* 149–154.

[71] Norinder, U. Single and Domain Made Variable Selection in 3D QSAR Applications. *J. Chemom.* **1996,** *10,* 95–105.

[72] Stanton, D. On the Physical Interpretation of QSAR Models. *J. Chem. Inf. Comput Sci.* **2003,** *43,* 1423–1433.

[73] Stanton, D. "personal communication", 2004.

[74] Guha, R.; Jurs, P. C. The Development of QSAR Models To Predict and Interpret the Biological Activity of Artemisinin Analogues. *J. Chem. Inf. Comput. Sci.* **2004,** *44,* 1440–1449.

[75] Svetnik, V.; Liaw, A.; Tong, C.; Culberson, C.; Sheridan, R.; Feuston, B. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *J. Chem. Inf. Comput. Sci.* **2003,** *42,* 1947–1958.

[76] Breiman, L.; Friedman, J.; Olshen, R.; Stone, C. *Classification and Regresion Trees;* CRC Press: Boca Raton, FL, 1984.

[77] Kier, L. A Shape Index From Molecular Graphs. *Quant. Struct.-Act. Relat. Pharmacol.,Chem. Biol.* **1985,** *4,* 109–116.

[78] Kier, L. Shape Indexes for Orders One and Three From Molecular Graphs. *Quant. Struct.-Act. Relat. Pharmacol.,Chem. Biol.* **1986,** *5,* 1–7.

[79] Kier, L. Distinguishing Atom Differences in a Molecular Graph Index. *Quant. Struct.-Act. Relat. Pharmacol.,Chem. Biol.* **1986,** *5,* 7–12.

[80] Randic, M.; Brissey, G.; Spencer, R.; Wilkins, C. Search for All Self-Avoiding Paths Graphs for Molecular Graphs. *Comput. Chem.* **1979,** *3,* 5–14.

[81] Wiener, H. Structural Determination of Paraffin Boiling Points. *J. Am. Chem. Soc.* **1947,** *69,*.

[82] Balaban, A. Higly Discriminating Distance Based Topological Index. *Chem. Phys. Lett.* **1982,** *89,* 399–404.

[83] Kier, L.; Hall, L. *Molecular Connectivity in Chemistry and Drug Research;* Academic Press: New York, 1976.

[84] Kier, L.; Hall, L. *Molecular Structure Description. The Electrotopological State;* Academic Press: London, 1999.

[85] Kier, L.; Hall, L. An Electrotopological-State Index for Atoms In Molecules. *Pharm. Res.* **1990,** *7,* 801–807.

[86] Gupta, S.; Singh, M.; Madan, A. Superpendentic Index: A Novel Topological Descriptor for Predicting Biological Activity. *J. Chem. Inf. Comput. Sci.* **1999,** *39,* 272–277.

[87] Katritzky, A.; Mu, L.; Lobanov, V.; Karelson, M. Correlation of Boiling Points with Molecular Structure. 1. A Training Set of 298 Diverse Organics and a Test Set of 9 Simple Inorganics. *J. Phys. Chem.* **1996,** *100,* 10400–10407.

[88] Vogel, A. *Textbook of Organic Chemistry;* Chaucer Press: London, 1970.

[89] Miller, K.; Savchik, J. A New Empirical Method to Calculate Average Emiprical Polarizabilities. *J. Am. Chem. Soc.* **1979,** *101,* 7206–7213.

[90] Mattioni, B. E. *The Development of Quantitative Structure-Activity Relationship Mode Physical Property and Biological Activity Prediction of Organic Compounds,* PhD thesis, Department of Chemistry, Pennsylvania State University, 2003.