Chapter 4

Generation of QSAR Sets Using a Self-Organizing Map

4.1 Introduction

As mentioned in Chapter 2, self-organizing maps[1] (SOM) are a class of unsupervised neural networks whose characteristic feature is their ability to map nonlinear relations in multi-dimensional datasets into easily visualizable two-dimensional grids of neurons. SOM's are also referred to as self-organized topological feature maps since the basic function of a SOM is to display the topology of a dataset, that is, the relationships between members of the set. SOM's were first developed by Kohonen in the 1980's, and since then they have been used as pattern recognition and classification tools in various fields including robotics,[2] astronomy,[3] and chemistry.

Neural networks, in general, have been used extensively in chemistry[4] and chemometrics and examples of applications in chemistry include spectroscopy,[5–8] prediction of NMR properties[9] and prediction of reaction products[10,11]

SOM's have also been applied to studies in the field of QSAR/QSPR.[12] The fundamental premise of QSAR studies is that structurally related (similar) compounds will have similar properties. Determining similarity is a complex task, and many methods exist such as principal components analysis and hierarchical cluster analysis. The fact that a SOM is able to extract topological information from a dataset makes it a valuable tool for detecting similarities in a dataset. Thus, it is to be expected that neighboring neurons in a two-dimensional SOM grid will be similar to each other. If each neuron in such a SOM grid can be assigned a molecule, groups of similar molecules can be identified.

Many studies have used a SOM to perform the actual QSAR[13–16] analysis by detecting relationships between structures and activities of interest. Other applications use SOM's at different stages of the QSAR study, for example, the use of a SOM to choose the best subset of molecular descriptors[17,18] to perform a QSAR analysis. However, another important step in QSAR study is the generation of training, cross-validation,

This work was published as Guha, R.; Serra, J.; Jurs, P.C., "Generation of QSAR Sets with a Self-Organizing Map", *J. Mol. Model. Graph.*, **2004**, *23*, 1–14.

and prediction sets. A number of methods exist, including random selection, activity-ranked binning, and sphere exclusion algorithms.[19] A number of studies have focused on approaches to select the training set. These approaches include classical statistical design methods such as Kennard-Stone,[8,20] and D-Optimal[8] as well as Kohonen neural networks.[8,20–22] Set selection is also an important step in QSAR modeling of chemical libraries. Most strategies for this are based on a combination of principal components analysis (PCA) for dimensionality reduction followed by statistical molecular design[23–25] (SMD).

The goal of this study was to implement a set generation technique, utilizing a SOM, together with whole molecule descriptors, to initially classify the dataset and subsequently use this classification to generate training, cross-validation, and prediction sets for QSAR studies, whose composition would mirror the overall composition of the entire dataset. The expectation was, that this technique should lead to the generation of QSAR models that exhibit equal or higher validity than models generated from subsets developed with random selection or activity-ranked binning. The distribution of the members of the training and prediction sets (with respect to each other in *descriptor space*) was also studied by calculating a molecular diversity index.[26] In addition, the results from SOM generated QSAR sets were compared to results obtained using QSAR sets created using traditional activity binning, as well as sets created using a sphere exclusion algorithm described by Golbraikh.[19]

4.2 Implementation of an SOM

A Kohonen self-organizing map (SOM) is an unsupervised neural network that uses as its inputs only the independent variables of the dataset here, molecular structure descriptors. The theoretical details of the SOM can be found in Section 2.2.2.

The implementation for this study consisted of a 13x13 grid. A Gaussian kernel was used, and the learning factor α was set to an initial value of 1.0 and was decremented with a constant decrement of 0.1 per training iteration. The dataset we used to test this method consisted of 333 molecules. According to Chen[11] the grid should contain 333 to 999 neurons. This translates to grid sizes ranging from 18x18 up to 31x31. However, we noted that for grids larger than 15x15 the SOM converged to a configuration in which the training set was mapped relatively evenly over the grid with little apparent clustering. In addition the use of larger grids increased the running times significantly. The method of choosing a grid size does appear to be arbitrary. However, given the fact that following

Chen's rule of thumb produced grids with hardly any clustering observable, we felt that examining smaller grid sizes was justified. Using a 13x13 grid of neurons the SOM usually required 80 to 90 training iterations for the grid neurons to converge to their final values. Depending on the number of descriptors used to represent each compound, this took approximately 3 to 6 minutes on an AMD 750MHz Duron processor running RedHat Linux 7.3.

After the SOM was trained, the results were analyzed to detect clusters of neurons. In this context, a cluster refers to neurons that have similar Euclidean distances from each other. As mentioned in Satoh,[10] "recognition of boundaries of clusters in a Kohonen network is a difficult task". This was implemented by considering two neurons, having a distance less than a user specified value, to be a part of the same cluster. Starting with an arbitrary neuron, we assigned an arbitrary class label. Next, we considered the distances to all the nearest neighbor neurons. Using the rule mentioned above, the neighboring neurons were assigned classes; either the same class as the initial neuron or the opposite class. This procedure was then repeated with all the neurons in the grid. An example of the grid layout after cluster detection (using three different threshold values) is shown in Fig. 4.1. The diagrams are based on the grid generated using the BCUT & 2D Autocorrelation descriptor combination, which are described in a subsequent section.

The final step in this procedure was to assign classes to the actual dataset members by submitting each dataset vector to the trained grid. The class of the closest grid neuron (in terms of Euclidean distance) was assigned to the dataset member.

The result of the cluster detection procedure was to divide the dataset into two classes. Fig. 4.2 shows how the classified dataset is distributed over the SOM. As mentioned before, the arbitrariness of cluster detection lies in the fact that the user must specify a distance threshold value. Too small a value or too large a value results in all the dataset members being assigned to the same class. As the threshold value progresses from zero to larger values the SOM generates a bulk class containing the majority of the dataset members and a minor class. At one point, the populations of both classes will be approximately equal, and then with further increase of the threshold value the populations once again get skewed. It is thus clear that the threshold value must be chosen carefully. Below we describe the method that we employed to arrive at a threshold value.

It should be noted that the classification of the dataset by the SOM is not intended to correspond to a classification based on any structure-activity relationship. The aim of the classification is to simply divide the dataset into two sets differing in structural features, as characterized by whole molecule descriptors.

4.3 Using the SOM to Create Sets

In this study, the SOM was used to generate training, prediction and cross-validation sets (hereafter referred to collectively as QSAR sets) for QSAR studies using the ADAPT[27,28] methodology. Previously, these sets had been generated by randomly selecting the requisite number of molecules from the binned (based on activity) dataset. However, owing to the random selection process, the binning procedure does not necessarily create sets that represent the composition of the whole dataset. Yan and Gasteiger[22] used a SOM to select QSAR sets, in which sets were created by simple selection of grid points. As a result thier method is similar to the sphere exclusion technique, in that there is a correspondence between the training and prediction set points in descriptor space. However the technique described by Yan and Gasteiger[22] does not necessarily maintain a correspondence between the composition of the QSAR sets and the overall dataset. Our method emphasizes the use of characteristic features of the dataset to create sets whose composition would mirror the overall dataset. This is achieved by using the SOM to divide the dataset into two classes, based on the molecular structure descriptors representing the compounds of the dataset. These two classes thus represent the SOM classification of the whole dataset into a major and minor class (say, Class I and Class II, respectively).

As described above, the threshold value controls the population of the two classes. We initially ran the SOM with the threshold value set to zero. The output of this run reported the distances between all the neurons in the grid. This distance information was used to determine the range of threshold values to be considered in subsequent runs of the SOM. The next step was to run the SOM several times in succession, with threshold values ranging from about 5% to 90% of the maximum distance reported in the initial run. Each run generated a set of class assignments. We considered those runs that generated a bulk class having approximately 80% of the entire dataset. The difference between the populations of the bulk and minor class for each of these runs, D, was noted. A large jump in the value of D was usually seen at one point in the series. This can be seen in Fig. 4.3, which plots D versus the threshold value (represented as a percentage of the maximum distance in the grid when the threshold value is set to zero). The descriptor subset supplied to these SOM runs was the MoRSE-WHIM subset. The classification results from the run that generated the lower value of D for the jump were used for the subsequent creation of QSAR sets. From Fig. 4.3 it is apparent that there is a large jump from 23% to 24% as well from 4% to 5%. However, we did not consider

these jumps since the number of molecules in the bulk class for these jumps was not close to 80% of the whole dataset. Instead, the grid configuration that corresponds to the jump from 9% to 11% had a bulk class that contained 80.1% of the whole dataset. Thus the grid results from the run using a threshold value of 11% were used subsequently. After the dataset had been classified, the information produced was used to create the actual QSAR sets. At this point the SOM had classified the dataset into two classes (Class I and Class II), members of each class being similar to each other but dissimilar to members of the other class.

Now, for example, say that Class I contains 75% of the whole dataset and Class II contains the other 25%. Our premise is that QSAR sets that contain Class I and Class II molecules distributed according to their percentages in the overall dataset will be more representative of the overall dataset and thus should lead to good predictive models. Continuing with the example, let us assume that we have a dataset of 100 molecules and the SOM classifier splits this dataset in to 75 molecules in class I and 25 molesules in class II. We also assume that for the QSAR sets, the training set should contain 80% of the dataset and the cross-validation and prediction sets should each contain 10%. To make the training set composition similar to that of the overall dataset it will have 80 compounds, of which 75% (60 compounds) will be from class I and 25% (20 compounds) will be from class II. Similarly the cross-validation and prediction sets will each have 10 compounds, of which 75% (8 compounds) will be from class I and 25% (2 compounds) will be from class II. Due to rounding, the final QSAR sets may not have the exact number of compounds described, but can differ by 1. The breakup of the QSAR sets among the SOM classes discussed above is represented diagrammatically in Fig. 4.4 with the exact numbers of compounds rounded appropriately.

Unlike methods such as the sphere exclusion method, discussed below, there is no guarantee that the QSAR sets generated cover the entire descriptor space. Though it is possible that a specific QSAR set is generated by sampling points from a small region of the grid, while still covering both classes, it appears that this does not occur. Fig. 4.5 shows the distribution of the QSAR sets over the grid. As can be seen, the members of each set seem to be relatively evenly distributed over the grid. The diagrams in Fig. 4.5 are based on the BCUT & 2D-Autocorrelation descriptor combination. The other QSAR sets generated from other Dragon[29] descriptor combinations investigated generated similar plots.

4.4 Sphere Exclusion

This method, described by Golbraikh,[19] uses the concept of molecular diversity[26] coupled with a sphere exclusion algorithm to generate training and prediction sets which satisfy the following criteria: points in the training and prediction sets should be close (in terms of descriptor space) to each other, and the training set should be diverse, as measured by the value of its diversity index.[26]

Golbraikh describes three types of sphere exclusion algorithms. A brief summary of the general sphere exclusion algorithm follows. For a training set with N compounds and described by K descriptors, the compound with the highest activity is first selected and placed in the training set. Next, a radius, R, is calculated. R is given by the formula

$$R = c\left(\frac{V}{N}\right)^{1/K} \tag{4.1}$$

where V is the volume of the space occupied by the points of the dataset in the descriptor space and c is a user defined constant termed the Dissimilarity Level (DL)[26] and essentially controls the number of molecules placed in the training and prediction sets. To simplify calculations, the descriptor space is normalized using the formula

$$X_{ij}^{n} = \frac{X_{ij} - X_{j,\min}}{X_{j,\max} - X_{j,\min}} \tag{4.2}$$

where X_{ij} is the non-normalized j'th descriptor for the i'th molecule and X_{ij}^{n} is the normalized value of the descriptor. Thus after normalization, $V = 1$ and the equation for the radius simplifies to

$$R = c\left(\frac{1}{N}\right)^{1/K} \tag{4.3}$$

After a value of R is obtained, a sphere with this radius is centered at the point chosen above, and all compounds that lie within this sphere (except the center point) are included in the prediction set and removed from the dataset so as not to be considered later. At this point if there are no more points left to consider the algorithm halts, otherwise the distances from the remaining points to the centers of all the spheres considered so far are calculated. The distance is given by

$$d_{ij} = \sqrt{\sum_{a=1}^{K}(X_{ia} - X_{ja})^2} \tag{4.4}$$

where X_i and X_j are the descriptor vectors for the i'th and j'th molecules respectively and K is the number of descriptors. One of the points is chosen to be the center of the next sphere and this process is repeated. The manner of choosing the next point gives rise to 3 variations of the sphere exclusion algorithm: the point that had the smallest d_{ij} , the point that had the largest d_{ij}, or randomly choosing a point. In this study we implemented the first option. The result of this algorithm is to generate a training and prediction set. Since the ADAPT methodology requires the use of a cross-validation set, we randomly selected the required number of molecules out of the training set to create the cross-validation set.

4.5 Descriptors for the SOM

The SOM requires that each compound be represented by a set of molecular structure descriptors. We used an external set of descriptors (from the Dragon[29] program), as opposed to the ADAPT descriptors since we wanted to classify the dataset in terms of global features, rather than specific structural trends. As a result, various subsets of Dragon descriptors which are holistic in nature were used, rather than ADAPT descriptors, many of which concentrate on specific structural features. Another reason for not using ADAPT descriptors is that the resultant QSAR sets would indirectly contain the information generated by the ADAPT descriptors and thus using same descriptors again during model development would lead to the possibility of biased models (in that the same information that was used to arrange the molecules would be used again when predicting their activity).

This technique, thus, proceeds in two stages and requires two sets of descriptors, preferably orthogonal in nature. In the first stage, one set of descriptors is used to classify the dataset with the SOM leading to creation of training, cross-validation and prediction sets. The second stage involves the generation of the actual QSAR model using the second set of ADAPT descriptors and the training, cross-validation and prediction sets created in the first stage.

As mentioned above the, descriptors for the first stage were taken from the Dragon program. Several combinations of the Dragon descriptors were selected to see if they could provide a holistic description of the molecules. The number of descriptors in each combination was reduced using correlation and identical testing before using them in the SOM algorithm. A brief description of the descriptors used for the SOM clustering follows. The size of each reduced Dragon descriptor set is shown in Table 4.1.

The BCUT metrics[30–32] are hybrid descriptors derived from the Burden parameters[30] which originally combined the atomic number of an atom and the bond types for adjacent and non-adjacent atoms. The BCUT metrics improve upon the number and type of atomic features that can be encoded. This descriptor has shown significant utility in the measurement of molecular diversity.[33]

Autocorrelation descriptors are based on the autocorrelation function defined[34] as

$$AC_l = \int_a^b f(x)f(x+l)\ dx \tag{4.5}$$

where $f(x)$ is a function of x, l is an interval of x and a and b are the limits of the interval under consideration. $f(x)$ is generally a time-dependent function or in the case of molecular descriptors a spatially-dependent function, in which the atoms of a molecule define the points in space and $f(x_i)$ represents some atomic property for the i^{th} atom. Both 2-D and 3-D autocorrelation descriptors can be calculated and for this study we restricted ourselves to three types of 2-D autocorrelation descriptors - Moreau-Broto,[35–37] Moran,[38] and Geary.[39] These descriptors sum products of atom properties of terminal atoms of all paths of a specific path length.

The Galvez topological charge indices[40–42] use the distance matrix to evaluate charge terms which characterize the charge transfer between individual atoms in the molecule.

The GETAWAY[43–45] descriptors are based on the information contained within the molecular influence matrix.[44] They combine the geometrical information in the influence matrix and topological information in the molecular graph weighted by various atomic properties. As a result, there are two sets of GETAWAY descriptors, the H & R GETAWAY, both of which we chose to use in the classification stage.

3D-MoRSE[46,47] descriptors are fixed length representations of 3D molecular structure and are based on the algorithm used for the analysis of electron diffraction data. Individual descriptors are obtained by considering different weighting functions as described in the literature.

WHIM[48–53] descriptors describe a molecule in terms of size, shape, symmetry and atom distribution and are based on a principal components analysis on the centered molecular coordinates with different weighting schemes.[52]

Our studies used both individual sets as well as combinations of the above descriptors. The sphere exclusion method also used the same combinations of external Dragon descriptors for the generation of the QSAR sets.

4.6 Results and Discussion

To test this method we generated QSAR models using the 333-compound pcD-HFR dataset that was studied by Mattioni.[54] The structures and activity values for all the molecules are contained in above-mentioned reference. To generate the QSAR sets we fed combinations of Dragon descriptors to the SOM, and its output was used to generate the sets. For each Dragon descriptor subset the sizes of the training, cross-validation and prediction sets were the same. To ensure a large enough training set, 80% of the dataset was placed in the training set and the remaining 20% was divided equally amongst the CV and prediction sets. The actual number of molecules in each set is summarized in Table 4.2 . After the QSAR sets were generated, we calculated ADAPT descriptors for the entire dataset of 333 molecules. This generated 248 descriptors for each molecule. The number of descriptors was then reduced via objective feature selection to generate a reduced pool of 74 descriptors. The reduced pool of ADAPT descriptors was then used with the QSAR sets created from each of the Dragon descriptor combination, to build nonlinear computational neural network (CNN) models using the ADAPT methodology. In total we used six combinations of Dragon descriptors to generate six nonlinear CNN QSAR models (Table 4.1).

4.6.1 Nonlinear CNN Models

To generate nonlinear models, the descriptor subsets selected by a genetic algorithm were fed to a 3-layer, fully-connected, feed-forward neural network to test fitness. The best neural network models were those that minimized the cost function defined as

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

where RMSE_{TSET} and RMSE_{CVSET} are the RMS errors for the training and cross validation sets respectively. Futher details of the development of CNN models using a genetic algorithm for model selection can be found in Section 3.4.2.

After several of the best (low cost) models were obtained a more rigorous analysis was performed on each model to identify the optimal neural network parameters. The results for the nonlinear models are summarized in Table 4.3. Though the number of descriptors in the two best models (see Table 4.4) is significantly lower than the number in the published model, they are of similar types, the majority being simple structural counts and topological path descriptors.

The MoRSE-2D Autocorrelation Dragon descriptor combination generated QSAR sets which produced a CNN model whose prediction set RMSE value was slightly larger than the original value whereas the prediction set error for the models that were generated from QSAR sets produced by using the GETAWAY and the MoRSE-WHIM Dragon descriptor sets match the predicted value. However, in all cases the RMSE's for the training and cross-validation set were significantly larger than those for the reported model. The higher cross-validation set error could indicate a loss of generalizability in these models. On the other hand the RMSE values for the training and cross-validation sets generated from the MoRSE-GETAWAY combination are much closer to those reported, though the prediction set error is now significantly larger. However, the attractive feature of the models generated from QSAR sets produced by MoRSE-2D Autocorrelation, GETAWAY and MoRSE-WHIM Dragon descriptor sets are that they are 5- or 6-descriptor models. Furthermore, the number of neurons in the hidden layers in these three models are all less than in the published model, indicating a simpler neural network.

Table 4.4 lists the descriptors present in the two best nonlinear models. The two best models have a similar set of ADAPT descriptors when compared to the published model, though none of them include a geometric descriptor. The R^2 values for the two best models (i.e., the models using QSAR sets generated using the MoRSE-2D Autocorrelation and MoRSE-WHIM Dragon descriptor combinations) are close to those reported for the best model. These are summarized in Table 4.5. The R^2 values for the training and cross-validation sets produced by the MoRSE-WHIM combination compare favorably to those published. The R^2 value for prediction set produced by the MoRSE-WHIM combination is a little higher than the reported value, but is not significantly larger. Considering the fact that the R^2 for the prediction set, produced by the MoRSE-2D Autocorrelation combination, is the same as that published this could indicate that a combination of MoRSE, 2D Autocorrelation and WHIM descriptor sets would lead to QSAR sets which would lead to a CNN model with better correlation coefficients overall.

However, it should be noted that though the R^2 value is a good test for evenly distributed data, it is not always reliable for an unevenly distributed dataset as the one used in this study. As a result we feel that the RMSE values provide a more reliable indication of the fitness of a model.

The plot for the predicted versus experimental values for the model generated using QSAR sets produced using the MoRSE-WHIM Dragon descriptor combination is shown in Fig. 4.6. Molecules in the prediction set were classified as outliers if their predicted value was two standard deviations away from the mean. This criterion led to

one outlier, whose structure is shown in Fig. 4.7. The best, published model also classified a single outlier. Ideally, we would like the same outliers to be detected by either method. Although this is not the case, it should be noted that there is a structural similarity in the outliers presented in Fig. 4.7. Although the original work does not provide an explanation of why that outlier is not predicted well, the fact that the SOM based technique predicts a structurally similar outlier indicates that that this technique is able to take into account similarity features of the dataset in the creation of the QSAR sets.

An important feature of the two best CNN models (using QSAR sets generated from the MoRSE-2D Autocorrelation and the MoRSE-WHIM Dragon descriptor combinations) is the consistency between the RMSE's for the training, cross-validation, and prediction sets. In many cases a low RMSE for the prediction set would be indicative of a good predictive model. However, at the same time, if the RMSE for the training and cross-validation sets are much lower than that of the prediction set it could indicate that the model lacks generalizability. Thus one would strive for models that have similar or consistent RMSE values for all the three QSAR sets. As can be seen, this does not occur for the original published results. However, for the best CNN models generated by the SOM-based method, though the RMSE's for the training and cross-validation sets are higher than those reported in the original model, the RMSE's are more consistent over all the three sets. The standard deviation of the RMSE's for the three QSAR sets in the original model is 0.11 whereas the standard deviations in the case of the two best models noted above are 0.02 in both cases. This suggests that the models generated by this method have both sufficient generalizability as well as predictive ability. However, apart from the conclusions regarding the nature of the models themselves, these results are indicative of the fact that the QSAR sets that are generated by the SOM are indeed similar to each other and representative of the data set as a whole thus leading to similar predictions made during training and after training (using the external prediction set).

We also reran the original, published 10–6–1 CNN model five times with different QSAR sets generated using activity binning. The results obtained are summarized in Table 4.6. As can be seen there is a large variation in the RMSE's for the three QSAR sets in each run. Furthermore, when compared to the RMSE's for the best CNN models generated using QSAR sets created by the SOM, we see that the SOM results in general lie midway between the RMSE values from the 10–6–1 models using QSAR sets from activity binning. We believe that this is a good indication for the consistency of results obtained using the SOM to generate representative QSAR sets.

It thus appears that the technique of using a group of external descriptors coupled with a SOM to generate sets for QSAR modeling do generate improved results. The ability of the SOM to detect similarities in the dataset allows us to generate sets that are more representative of the overall data set. As a result models with fewer parameters (i.e., descriptors) are able to produce results comparable to the original model that had nearly twice the number of parameters and in addition produce consistent RMSE's over the three QSAR sets.

4.6.2 Sphere Exclusion

For comparison, results of CNN models generated using different QSAR sets created by the sphere exclusion method are presented in Table 4.7. For each set of external descriptors used, the model with lowest cost is reported. None of the models seem to be significantly better than the published model. The architectures are not significantly simpler than the reported 10–6–1 architecture and the R^2 values and RMSE's are comparable, though none of the models seem to provide an improvement over the published statistics. In addition, there is not much of a difference in the RMSE values for models that are generated from QSAR sets that were created using different Dragon descriptor combinations. However, when comparing the results from the sphere exclusion method to those obtained from the SOM technique it appears that the SOM generated QSAR sets produce better models in terms of size (i.e., requiring fewer descriptors), with RMSE's being comparable. In addition the RMSE's for the three QSAR sets in the models generated by the sphere exclusion method do not show much consistency. The RMSE for the prediction set is usually higher than the RMSE's for the training and cross-validations sets by 0.1 to 0.3. This is similar to the nature of the RMSE's in the original model. Due to the nature of the sphere exclusion algorithm one would expect that the resultant QSAR sets would be similar to each other and thus lead to consistent RMSE's. The fact that it does not, is a possible indication that a simple Euclidean distance between individual molecular descriptor vectors is not sufficient to characterize similarity of molecules of in a dataset. Thus the sphere exclusion method does not appear to generate QSAR sets that can produce models with both generalizability as well as predictive ability for this dataset.

4.6.3 Randomization Studies

The best nonlinear model (i.e., the one generated using QSAR sets produced by the MoRSE-WHIM Dragon descriptor combination) was subjected to randomization tests. The first set of tests involved generating random training, cross-validation and prediction sets. These sets were then used to generate a nonlinear model with a 6–5–1 CNN architecture five times (each time using randomly generated sets) and noting the average RMSE's. In addition, the variance between the five individual runs for the random sets was also compared to the variance for five runs of the original QSAR sets that gave the best 6-descriptor CNN model. The correlation coefficient for each of the sets in each of the runs was also compared to the correlation coefficients for the best model. The results are summarized in Table 4.8. The average correlation coefficient (R^2) for the random training, cross-validation and prediction sets were 0.75, 0.73, and 0.56 respectively. These values would indicate that the KSOM technique is not much better than random set generation. As mentioned above, R^2 is not always reliable for an unevenly distributed dataset such as the one used in this study and as a result we feel that the RMSE values provide a better indicator of the goodness of a model. Though the RMSE's for the training and cross-validation sets are comparable, the prediction set RMSE is much larger for the random sets. In addition, comparing the standard deviation in the RMSE values for the five runs for the random and KSOM sets indicates that the KSOM technique is more consistent compared to the random approach. For the original best model the standard deviations for the three sets were 0.005, 0.01 and 0.02, respectively. For the random sets the standard deviations were 0.02, 0.03 and 0.13 respectively, indicating that predictions made using the random sets were not consistent over several runs. Once again, we believe that this is evidence for the KSOM's ability to generate good sets based on features of the dataset.

The next randomization test consisted of regenerating the best nonlinear model (using the ADAPT descriptors as reported in Table 4.4) but scrambling the dependent variable. With the scrambled dependent variable, the best CNN model was regenerated using the original QSAR sets. This process was repeated five times, each time scrambling the dependent variable and the average RMSE and R^2 values for the training, cross-validation and prediction sets for the five runs were noted. It would be expected that the resultant model would have relatively high RMSE's for the three sets, as well as low R^2 values. This was indeed the case with the training, cross-validation and prediction sets having RMSE values of 1.04, 1.00 and 0.97 respectively (Table 4.9). In addition

the R^2 values for the three sets were 0.17, 0.09 and 0.01, respectively. Compared to the RMSE and R^2 values for the best model, it appears that chance correlations played little (if any) part in the results for the best model.

Finally a randomization test was carried out to investigate the role of chance correlations in the genetic algorithm (i.e., the descriptor selection algorithm). This was carried out by generating one hundred CNN models using a 6–5–1 architecture and the QSAR sets generated by the SOM (using the MoRSE-WHIM Dragon descriptor subset). However in each run, six ADAPT descriptors were randomly selected from the reduced pool. One would assume that the RMSE and R^2 values for the models generated by randomly selecting descriptors would be worse than for the best reported model but not as poor compared to the runs using a scrambled dependent variable. This can be explained by noting that since the dependent variable is not scrambled there will be some correlation with the descriptors selected. However due to the fact that we randomly select descriptors this correlation will not be as significant compared to descriptor selection using a genetic algorithm, which looks for descriptor subsets that are well correlated with the dependent variable and hence produce models with low cost functions. Thus this test ensures that the specific set of descriptors selected by the genetic algorithm did not arise by chance alone. The results for this test are provided in Table 4.10. As can be seen, the average RMSE for all three sets are higher than those reported for the best model, though the differences are not as significant compared to the results from the scrambled dependent variable test. The R^2 values are also lower than for the best reported model but are not as poor when compared to the results from the scrambled dependent variable test.

The results from the randomization tests described above thus indicate that chance correlations played little (if any) role in both the descriptor selection algorithm as well in the final model itself.

4.7 Diversity Indices and SOM Generated Sets

The SOM was used to prepare training and prediction sets such that they would be heterogeneous in nature and representative of the whole dataset. The molecular dataset diversity index[26] has been developed to quantify the diversity of a dataset and the correspondence between training and prediction sets. This metric provides a quantitative estimate of the similarity between the training and prediction sets. Golbraikh describes three quantities - $M_{(\text{test, train})}$, $M_{(\text{train,test})}$ and I_{train}. The quantity of interest here is

$M_{(test,train)}$, which measures the diversity of the training set with respect to the prediction set. The value of $M_{(test,train)}$ depends on both the algorithm used to generate sets as well as the distribution of the data set in the descriptor space. In general lower values of $M_{(test,train)}$ indicate that the points in the prediction set are closer (or correspond better) to the points in the training set. However the evaluation of $M_{(test,\ train)}$ depends on the value of an arbitrary value termed the Dissimilarity Level (DL). Golbraikh does not go into detail regarding the choice of a dissimilarity level. Hence, we calculated $M_{(test,\ train)}$ values at increasing DL values for each Dragon descriptor combination, plotted them (Fig. 4.8), and correlated the behavior of the plots with the CNN model statistics. One would expect that for training and prediction sets which correspond well with each other (i.e., a prediction set point corresponds to some training set point) the $M_{(test,train)}$ should rapidly fall to zero with increasing DL values. However, another view would be to consider the training and prediction sets to be well distributed throughout descriptor space of the dataset. In such a case the correspondence between the two sets would not necessarily be very good and one would observe higher values of $M_{(test,train)}$ for a given DL value. This might lead one to conclude that such a situation would lead to bad model statistics. However, Fig. 4.8 indicates otherwise. From the plot we see that the curves for the MoRSE-2D Autocorrelation and the MoRSE-WHIM combinations remain constant at an $M_{(test,train)}$ value of 1 for all DL values up to approximately 2 and the MoRSE-GETAWAY combination remains at 1 up to nearly 2.5. From Table 4.3 we see that the MoRSE-GETAWAY combination has the best training and cross-validation set errors of all the sets tested, but its prediction set error is higher. At the same time, the training and prediction set errors for the MoRSE-WHIM and MoRSE-2D Autocorrelation combinations are larger than for the MoRSE-GETAWAY combination - but their order follows the trend in the graph. Sets that remain at a $M_{(train,test)}$ value of 1 for higher DL values appear to lead to lower RMSE's for the training and cross-validation sets.

If one considers the prediction set errors, a similar trend is seen. Sets whose $M_{(test,train)}$ vs. DL plots remain at a $M_{(test,train)}$ value of 1 for larger values of DL appear to lead to better prediction set errors. However this should not be considered as an absolute as the plot for the GETAWAY set does not follow this trend. In fact the prediction set error is equal to that for the MoRSE-WHIM set., but the $M_{(train,test)}$ value drops below 1 for DL values of 0.7 onwards. Thus the values of $M_{(test,train)}$ for the GETAWAY set are lower for a given DL value, indicating a better correspondence between the training and prediction sets. However, though this leads to a good prediction set error, the training and cross-validation set errors are quite large. This could imply

that lower $M_{(test,train)}$ values might lead to better prediction set errors but at the same time would lead to a loss of generalizability as evidenced by the training and cross-validation set errors.

Though the use of an arbitrary DL value in the evaluation of $M_{(test,train)}$ values does make interpretation of $M_{(test,train)}$ values slightly ambiguous, we feel that the technique we describe does provide some indication as to whether a training set might lead to good training and prediction set errors, based on diversity index information.

4.8 Conclusions

This study used a Kohonen self-organizing map to investigate whether a similarity based set generation method would lead to better QSAR models. Multiple runs using different sets of Dragon descriptors were used to generate training, cross-validation, and prediction sets, which were in turn used to create QSAR models. The best model obtained by this method did improve upon the previously published model in terms of model size. Although the actual RMSE values were not significantly better than those published, they were consistent and exhibited a lower standard deviation over the three QSAR sets compared to the original results. QSAR sets were also generated using a sphere exclusion[19] technique. Models generated using these QSAR sets did not show any significant improvement in terms of statistics or model size over the published results. When compared to the models generated using QSAR sets created by the SOM, we noted that there was no significant improvement in the statistics of the models generated by the sphere exclusion methods. Furthermore, the RMSE's of the three QSAR sets generated by the sphere exclusion method were not as consistent as those generated by the SOM and exhibited standard deviations similar to those of the original QSAR sets obtained by activity binning. However, the SOM did lead to models that were significantly simpler than those generated using the sphere exclusion method or activity binning (the published results). Randomization tests indicated that the models generated did not arise due to chance correlations. The use of the $M_{(test,train)}$ diversity index provided an indication of the Dragon descriptor set's ability to generate good QSAR sets which in turn lead to QSAR models.

Though the study did lead to a better model than that published, it involved a number of arbitrary decisions such as the choice of initial descriptors to submit to the SOM as well as choosing a specific SOM split out of several runs. The algorithm could be substantially improved by implementing a method to optimize the threshold value so

that classification of molecules in the SOM could be automated. Another improvement would be in the choice of initial descriptors. Since this study was exploratory in nature, we restricted ourselves to certain subsets of Dragon descriptors which we deemed to be holistic in nature. That is, we chose Dragon descriptor sets that appeared to characterize the whole molecule, rather than characterizing specific molecular features. In addition the choice of Dragon descriptors was also guided by the fact that we did not want to use ADAPT descriptors during the initial classification process. Clearly, there remains an element of arbitrariness in the selection of Dragon descriptor sets. This need not be the case and by including more or even all Dragon descriptors (followed by a PCA to obtain the main contributing components) the initial classification might be better. In addition though the evaluation of $M_{(\text{test,train})}$ does involve an arbitrary constant, it seems that looking at the trend rather than individual values (for fixed DL values) can be used to make a decision on which Dragon sets could be used for further study.

Table 4.1. Type and number of Dragon descriptors used by the SOM to generate training, cross-validation, and prediction sets for QSAR models.

Descriptor Name	No. of Descriptors	References
BCUT	123	30–32
BCUT & Galvez Topological Indices	63	30–32, 40–42
GETAWAY	128	43–45
MoRSE & 2D Auto Correlation	173	35–39, 46, 47
MoRSE & GETAWAY	223	43–47
MoRSE & WHIM	139	46–53

Table 4.2. Summary of the number of molecules present in the training, cross-validation and prediction sets. The sizes of these sets were the same for all the Dragon descriptor subsets investigated.

Set	Number of Molecules	Percentage of Molecules
Training	267	80.1
Cross-Validation	32	9.6
Prediction	34	10.3
Total	333	100

Table 4.3. Summary of the nonlinear CNN models using training, cross-validation, and prediction sets created by the SOM and Dragon descriptor combinations.

		RMSE			R^2		
Dragon Descriptor	CNN Arch.	TSET	CVSET	PSET	TSET	CVSET	PSET
BCUT - 2D Autocorrelation	5–3–1	0.63	0.68	0.79	0.68	0.60	0.67
BCUT - Galvez Topological Indices	5–3–1	0.62	0.62	0.71	0.69	0.66	0.64
GETAWAY	5–2–1	0.68	0.60	0.73	0.64	0.76	0.67
MoRSE - 2D Autocorrelation	5–3–1	0.63	0.63	0.68	0.68	0.60	0.74
MoRSE -GETAWAY	9–5–1	0.49	0.59	0.76	0.80	0.58	0.80
MoRSE - WHIM	6–5–1	0.60	0.61	0.65	0.75	0.78	0.64
Published Results[54]	10–6–1	0.45	0.49	0.66	0.84	0.78	0.64

Table 4.4. ADAPT descriptors present in the two best nonlinear CNN models.

MoRSE & 2D Autocorrelation			MoRSE & WHIM		
Descriptor	Type	Range	Descriptor	Type	Range
N7CH	Topo	7.0 - 28.0	V6P7	Topo	2.1 - 0.5
MOLC-8	Topo	0.6 - 2.8	WTPT-4	Topo	0.0 - 12.2
NDB-13	Topo	0.0 - 7.0	N7CH	Topo	7.0 - 28.0
NAB-15	Topo	6.0 - 23.0	NDB-13	Topo	0.0 - 7.0
WPSA-3	Hybrid	17 - 57.4	MDE-23	Topo	0.0 - 28.1
			RPCS	Hybrid	0.0 - 8.1

Topo indicates a topological descriptor. N7CH, number of seventh order chains index;[55–57] MOLC-8, average distance sum connectivity[58,59] (topological index J); NDB-13, number of double bonds; NAB-15, number of aromatic bonds; WPSA-3, partial positive surface area multiplied by the total molecular surface area divided by 1000;[60] RPCS, relative positive charged surface area;[60] MDE-23, molecular distance edge between primary and secondary carbons;[61] WTPT-4, sum of atom ID's for oxygens[62]

Table 4.5. Comparison of R^2 values for the training, cross-validation, and prediction sets created by the SOM using Dragon descriptors.*

	Training Set	Cross-validation Set	Prediction Set
MoRSE - 2D Autocorrelation	0.68	0.60	0.64
MoRSE - WHIM	0.75	0.78	0.67
Published[54]	0.83	0.78	0.64

* The models produced were CNN models. See Table 4.3 for the model architectures.

Table 4.6. A summary of the RMSE's for the 10–6–1 nonlinear CNN models using five QSAR sets generated by activity binning.

Serial No.	Training Set	Cross-validation Set	Prediction Set
1	0.45	0.59	0.81
2	0.45	0.52	0.73
3	0.44	0.63	0.95
4	0.64	0.64	1.00
5	0.67	0.61	0.95

Table 4.7. Summary of the best nonlinear CNN models generated from QSAR sets created using the sphere exclusion algorithm.

		RMSE			R^2		
Dragon Descriptor*	CNN Arch.	TSET	CVSET	PSET	TSET	CVSET	PSET
BCUT - 2D	9–3–1	0.55	0.54	0.87	0.74	0.78	0.33
Autocorrelation							
BCUT - Galvez	9–8–1	0.46	0.50	0.87	0.83	0.81	0.36
Topological Indices							
GETAWAY	8–5–1	0.56	0.56	0.63	0.75	0.80	0.67
MoRSE - 2D	9–8–1	0.49	0.53	0.68	0.81	0.82	0.68
Autocorrelation							
MoRSE -	8–6–1	0.52	0.58	0.64	0.79	0.84	0.67
GETAWAY							
MoRSE - WHIM	7–6–1	0.50	0.57	0.82	0.80	0.77	0.52
Published Results[54]	10–6–1	0.45	0.49	0.66	0.84	0.78	0.64

* The external descriptor set used by the sphere exclusion algorithm to create the training and prediction sets

Table 4.8. Comparison of statistics for training, cross-validation, and prediction sets generated randomly versus sets created by the SOM using the MoRSE-WHIM Dragon descriptor combination.*

	Random Sets			MoRSE - WHIM Sets		
	Mean RMSE	Std. dev.	Mean R^2	Mean RMSE	Std. dev	R^2
TSET	0.57	0.02	0.75	0.58	0.005	0.74
CVSET	0.59	0.03	0.73	0.57	0.010	0.76
PSET	0.80	0.13	0.56	0.63	0.020	0.63

* The statistics are from a nonlinear CNN model using a 6–5–1 architecture. The same descriptors were used in both models.

Table 4.9. RMSE values for a nonlinear CNN Model* using a scrambled dependent variable using training, cross-validation, and predictions sets created by the KSOM using the MoRSE-WHIM Dragon descriptor combination.

	Scrambled		Original	
	Mean RMSE	Mean R^2	Mean RMSE	R^2
TSET	1.04	0.17	0.58	0.74
CVSET	1.00	0.09	0.56	0.76
PSET	0.97	0.01	0.59	0.63

* The model was generated using a 6–5–1 CNN architecture and the ADAPT descriptors reported for the best nonlinear model.

Table 4.10. A summary of the RMS errors and R^2 values for one hundred runs of the best CNN architecture (6–5–1) using randomly selected ADAPT descriptors.*

	Mean RMSE	Std. Dev.	Mean R^2	Std. Dev.
Training Set	0.81	0.09	0.47	0.11
Cross-Validation Set	0.84	0.08	0.36	0.13
Prediction Set	0.84	0.09	0.28	0.13

* QSAR sets used in these models were created by the KSOM using the MoRSE-WHIM Dragon descriptor combination.

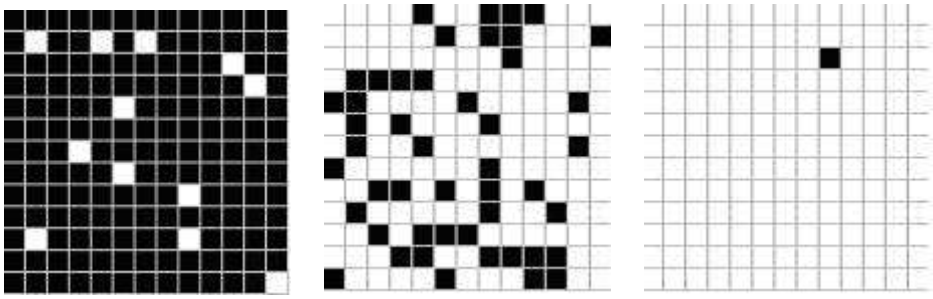


Fig. 4.1. A graphical representation of the SOM after the cluster detection step using the BCUT & 2D-Autocorrelation Dragon[29] descriptor subset. Black and white squares represent the individual classes. Grids A, B and C were obtained by setting the threshold value to 0, 1.2 and 3.6 respectively. Grid B was used to generate the final QSAR sets for this Dragon descriptor subset.

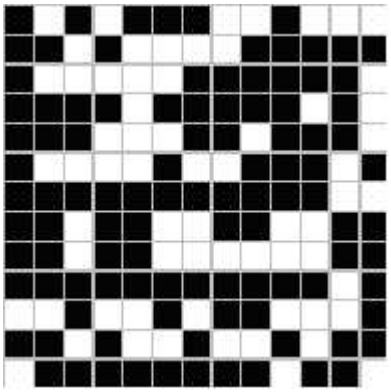


Fig. 4.2. A graphical representation of the distribution of whole dataset on the grid after it has been divided into two classes based on the BCUT & 2D-Autocorrelation Dragon[29] descriptor combination. Black and white squares represent the two different classes.

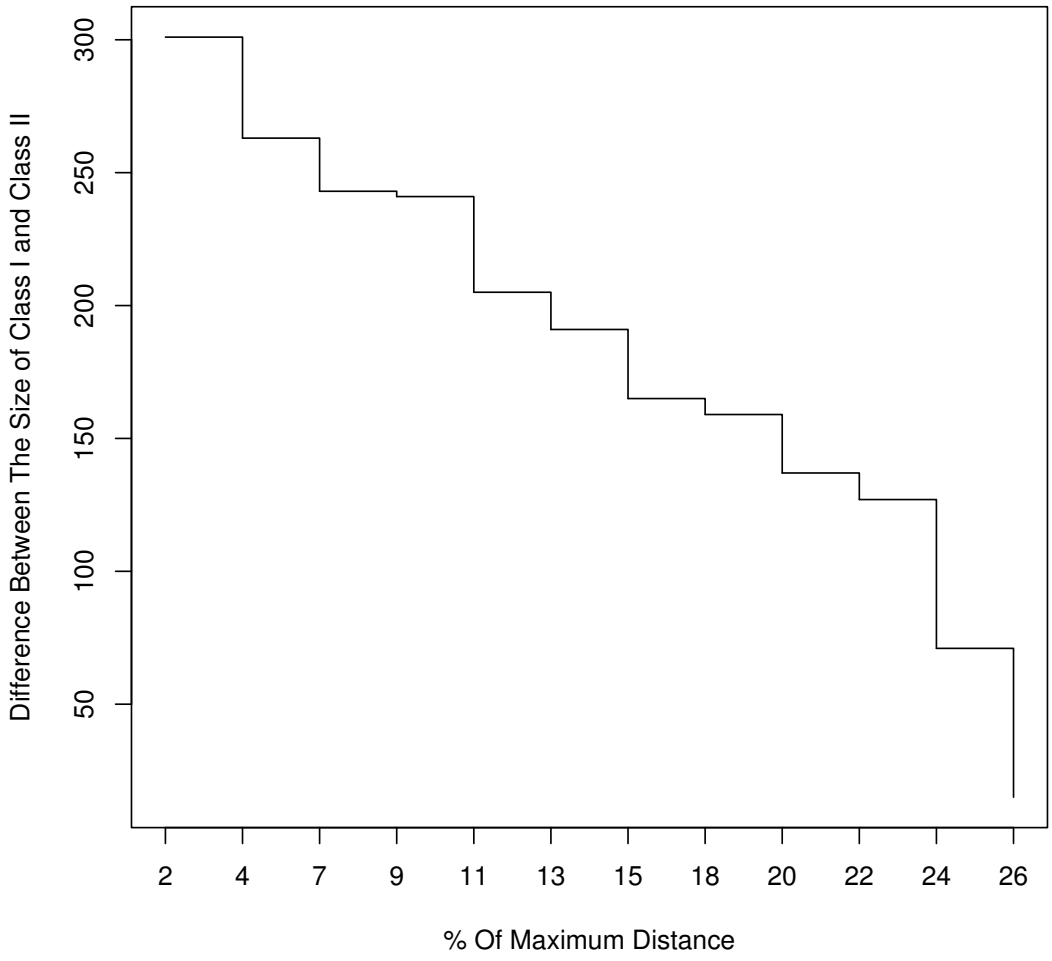


Fig. 4.3. A plot showing the variation of D (the difference in size between major and minor SOM classes) versus the threshold value for the SOM. In this plot the threshold value is represented as a percentage of the maximum distance in the grid for a SOM in which the threshold value was set to 0. The descriptor set used to generate the grids described in the plot was the MoRSE-WHIM Dragon[29] descriptor subset.

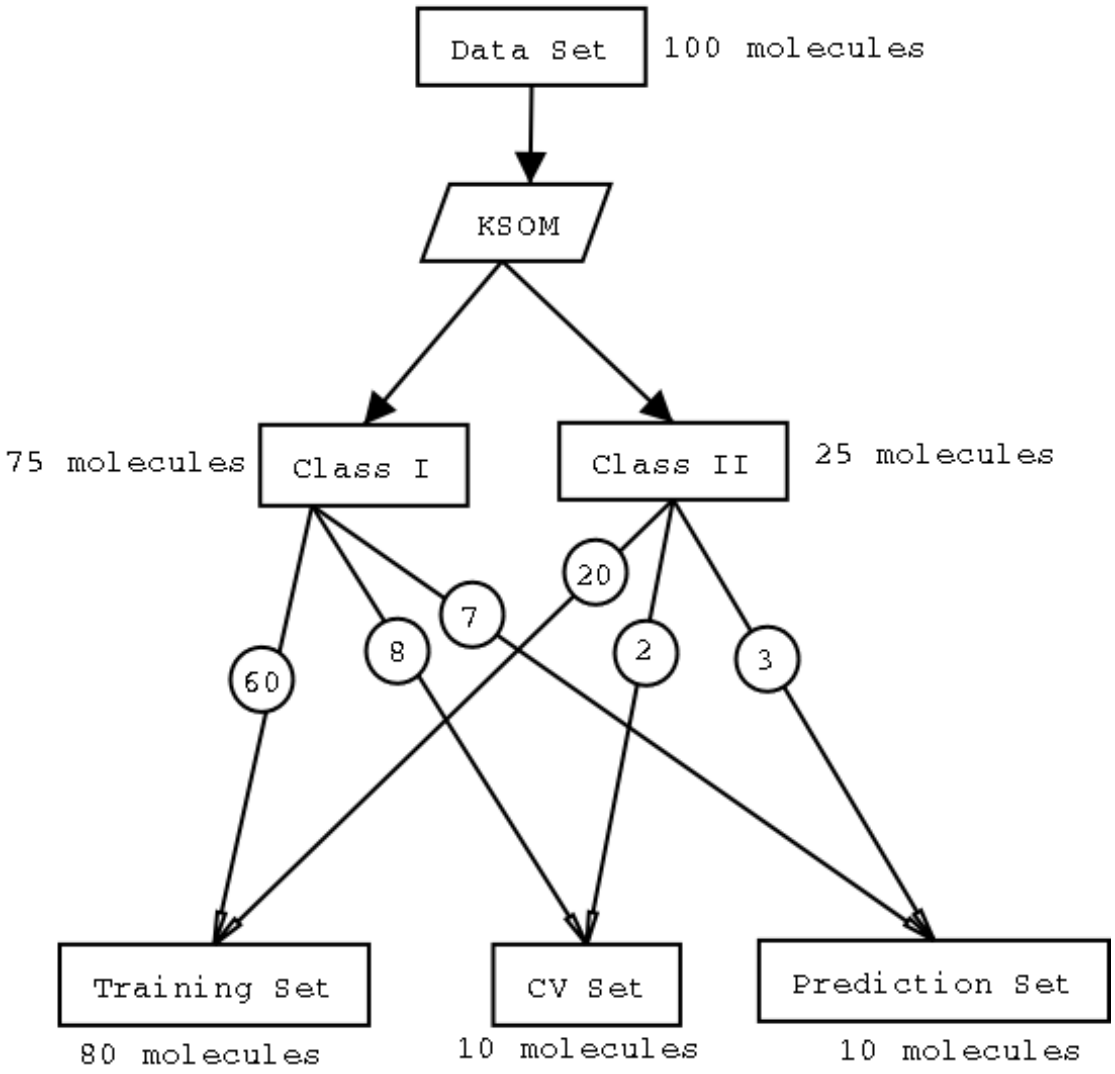


Fig. 4.4. A diagrammatic representation of the method we use to generate QSAR sets from the SOM classification of the whole dataset. The numbers within circles are the number of molecules from that class that present in the specific QSAR set.

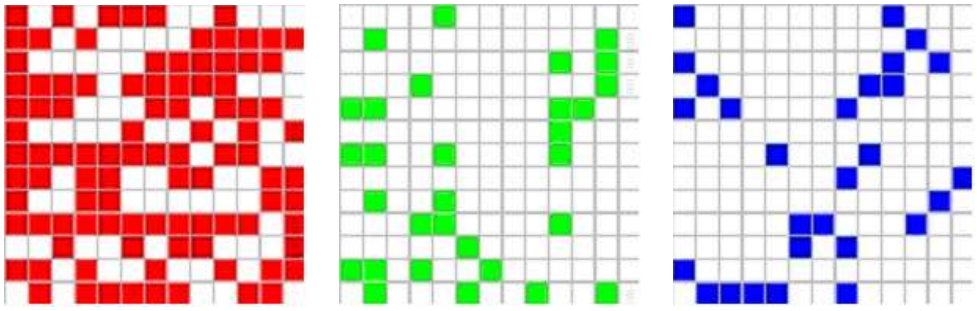


Fig. 4.5. The three diagrams represent the distribution of the QSAR sets over the surface of the SOM. The grid was trained with the BCUT & 2D-Autocorrelation Dragon[29] descriptor combination.

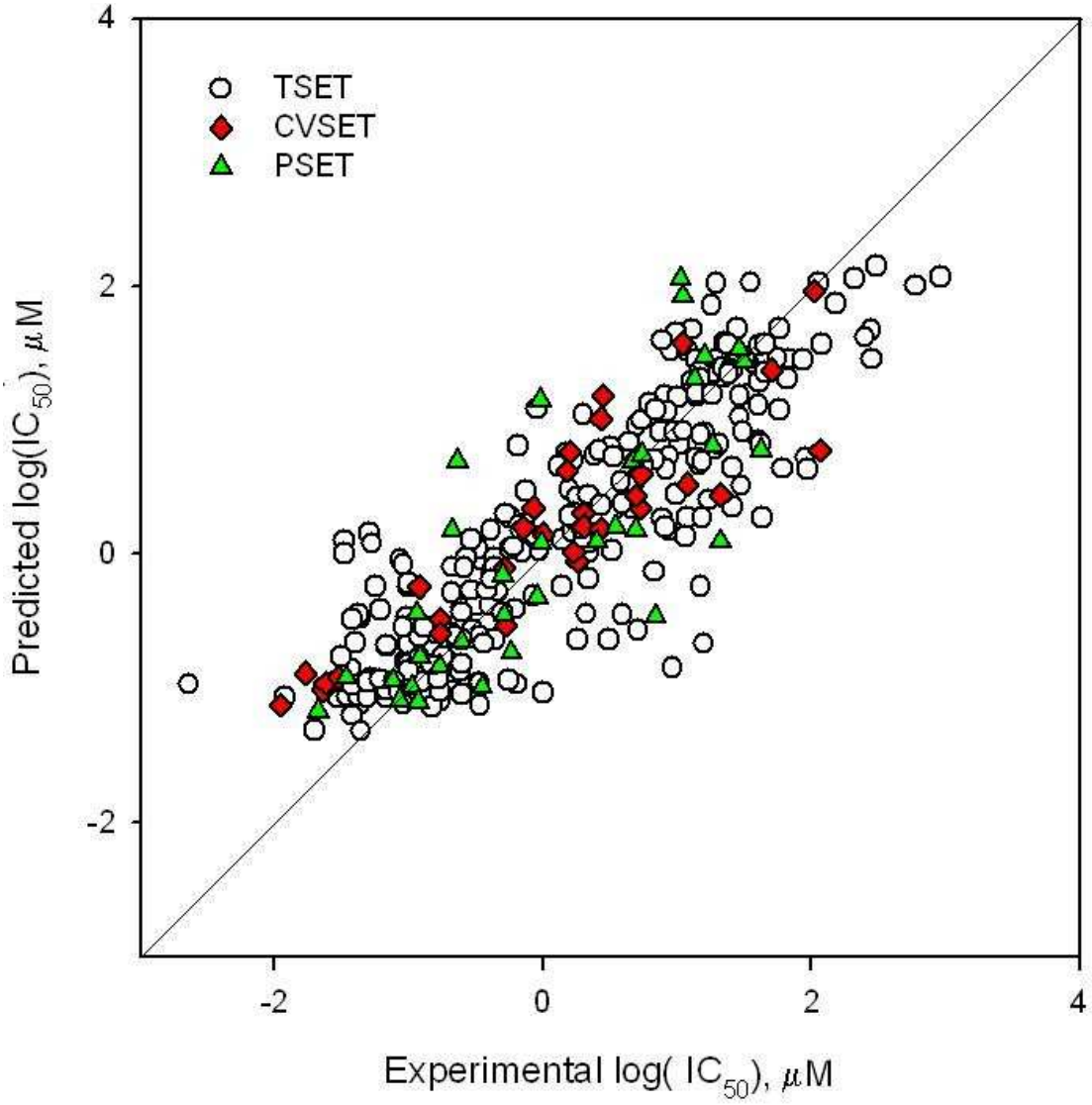
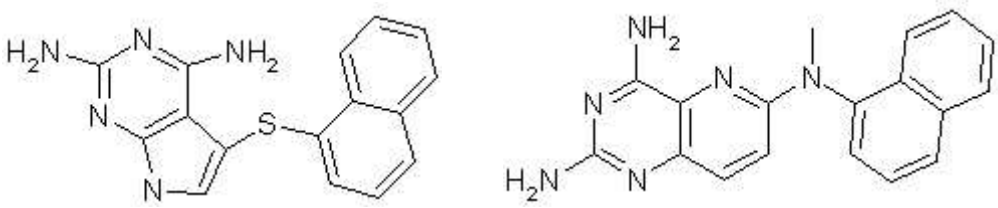


Fig. 4.6. Plot of experimental vs. predicted log IC_{50} for the 6–5–1 CNN model Generated Using Training, Cross-validation, and Prediction Sets Created Using the SOM and MoRSE - WHIM Dragon[29] Descriptor Combination.



Outlier detected by best CNN model in this study

Published outlier for the best pcDHFR model[54]

Fig. 4.7. Prediction set outliers.

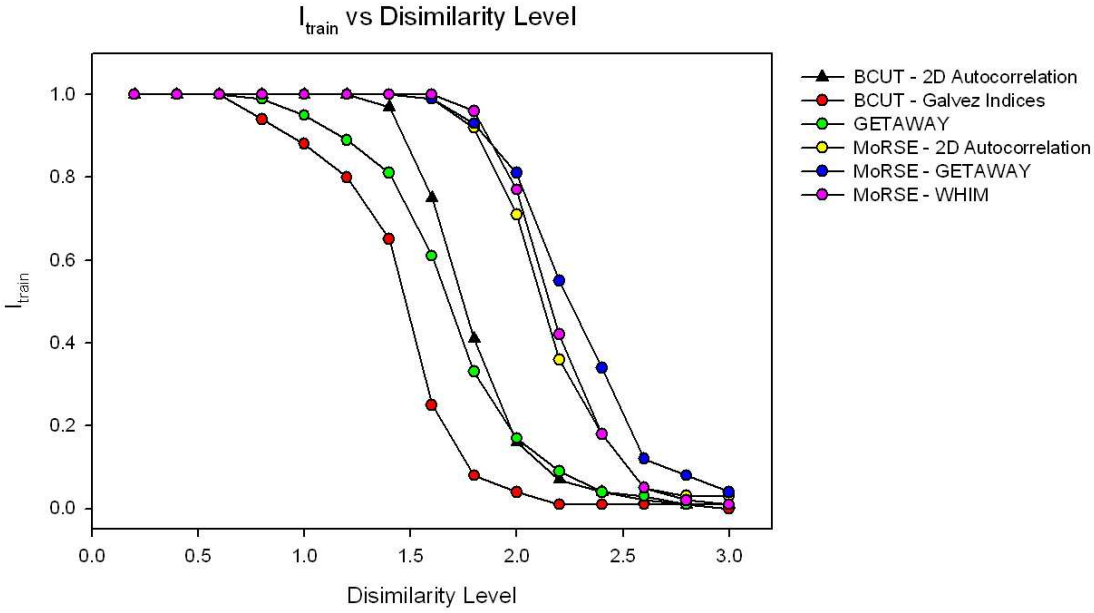


Fig. 4.8. Plot of dissimilarity level vs. $M_{(test,train)}$ for the various Dragon[29] sets studied.

References

[1] Kohonen, T. *Self Organizing Maps;* volume 30 of *Springer Series in Information Sciences* Springer: Berlin, 1994.

[2] Janet, J.; Gutierrez, R.; Chase, T.; White, M.; Sutton, J. Autonomous Mobile Robot Global Self Localization Using Kohonen and Region Feature Neural Networks. *Journal of Robotic Systems* **1997,** *14,* 263–282.

[3] Naim, A.; Ratnatunga, K.; Griffiths, R. Galaxy Morphology Without Classification: Self Organizing Maps. *Astrophysical Journal Supplement Series* **1997,** *111,* 357-367.

[4] Gasteiger, J.; Zupan, J. Neural Networks in Chemistry. *Angew. Chem. Int. Ed. Engl.* **1993,** *32,* 503–527.

[5] Daniel, N.; Lewis, I.; Griffiths, P. Interpretation of Raman Spectra of Nitro Containing Explosive Materials. Part II: the Implementation of Neural, Fuzzy and Statistical Models for Unsupervised Pattern Recognition.. *Appl. Spectr.* **1997,** *51,* 1854–1867.

[6] Vander Heyden, Y.; Vankeerbergghen, P.; Novic, M.; Zupan, J.; Massart, D. The Application of Kohonen Neural Netoworks to Diagnose Calibration Problems in Atomic Absorption Spectroscopy. *Talanta* **2000,** *51,* 455–466.

[7] Novic, M.; Zupan, J. Investigation of Infra-Red Spectra-Structure Correlation Using Kohonen and Counterpropagation Neural Networks. *J. Chem. Inf. Comput. Sci.* **1995,** *35,* 454-466.

[8] 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. *Chemometrics and Intelligent Laboratory Systems* **1996,** *33,* 35–46.

[9] Aires-de Sousa, J.; Hemmer, M. C.; Gasteiger, J. Prediction of 1H NMR Chemical Shifts Using Neural Networks. *Anal. Chem.* **2002,** *74,* 80–90.

[10] Satoh, H.; Sacher, O.; Nakata, T.; Chen, L.; Gasteiger, J.; Funatsu, K. Classification of Organic Reactions: Similarity of Reactions Based on the Electronic Features of Oxygen Atoms at the Reaction Sites. *J. Chem. Inf. Comput. Sci.* **1998,** *38,* 210–219.

[11] Chen, L.; Gasteiger, J. Knowledge Discovery in Reaction Databases: Landscaping Organic Reactions By a Self Organizing Neural Network. *J. Am. Chem. Soc.* **1997,** *119,* 4033–4042.

[12] Manallack, D.; Livingstone, D. Neural Networks in Drug Discovery: Have They Lived Upto Their Promise?. *Eur. J. Med. Chem.* **1999,** *34,* 195–208.

[13] Tetko, I.; Kovalishyn, V.; Livingstone, D. Volume Learning Algorithm Artificial Neural Networks for 3D QSAR Studies. *J. Med. Chem.* **2001,** *44,* 2411-2420.

[14] Bienfait, B. Applications of High Reolution Self Organizing Maps to Retrosynthetic and QSAR Analysis. *J. Chem. Inf. Comput. Sci.* **1994,** *34,* 890–898.

[15] Rose, V.; Croall, I.; Macfie, H. An Application of Unsupervised Neural Network Methodology Kohonen Topology-Preserving Mapping to QSAR Analysis. *Quant. Struct.-Act. Relat.* **1991,** *10,* 6–15.

[16] Anzali, S.; Barnickel, G.; Krug, M.; Sadowski, J.; Wagener, M.; Gasteiger, J.; Polanski, J. The Comparison of Geometric and Electronic Properties of Molecular Surfaces By Neural Networks: Application to the Analysis of Corticosteroid-Binding Globulin Activity of Steroids. *J. Comp. Aid. Molec. Des.* **1996,** *10,* 521–534.

[17] Gramatica, P.; Consonni, V.; Todeschini, R. QSAR Study of the Tropospheric Degradation of Organic Compounds. *Chemosphere* **1999,** *38,* 1371–1378.

[18] Espinosa, G.; Arenas, A.; Giralt, F. An Integrated SOM Fuzzy ARTMAP Neural System for the Evaluation of Toxicity. *J. Chem. Inf. Comput. Sci.* **2002,** *42,* 343–359.

[19] Golbraikh, A.; Tropsha, A. Predicitve QSAR Modeling Based on Diversity Sampling of Experimental Datasets for the Test and Training Set Selection. *J. Comp. Aid. Molec. Des.* **2002,** *16,* 356–369.

[20] Kocjancic, R.; Zupan, J. Modelling of the River Flowrate: The Influence of Training Set Selection.. *Chemometrics and Intelligent Laboratory Systems* **2000,** *54,* 21–34.

[21] Kirew, D.; Chretien, J.; Bernard, P.; Ros, F. Application of Kohonen Neural Networks in Classification of Biologically Active Compounds. *SAR and QSAR in Environmental Research* **1998,** *8,* 93.

[22] Yan, A.; Gasteiger, J. Prediction of Aqueous Solubility of Organic Compounds Based on 3D Structure Representation. *J. Chem. Inf. Comput. Sci.* **2003,** *43,* 429-434.

[23] Andersson, P.; Sjostrom, M.; Wold, S.; Lundstedt, T. Strategies for Subset Selection of Parts of an In House Chemical Library. *J. Chemom.* **2001,** *15,* 353–369.

[24] Linusson, A.; Gottfries, J.; Olsson, T.; Ornskov, E.; Folestad, S.; Norden, B.; Wold, S. Statistcal Molecular Design, Parallel Synthesis and Biological Evaluation of a Library of Thrombin Inhibitors. *J. Med. Chem.* **2001,** *44,* 3424–3439.

[25] Linusson, A.; Gottfries, J.; Lindgren, F.; Wold, S. Statistical Molecular Design of Building Blocks For Combinatorial Chemistry. *J. Med. Chem.* **2000,** *43,* 1320–1328.

[26] Golbraikh, A. Molecular Dataset Diversity Indices and Their Applications to Comparison of Chemical Databases and QSAR Analysis. *J. Chem. Inf. Comput. Sci.* **2000,** *40,* 414–425.

[27] 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.

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

[29] Todeschini, R.; Consonni, V.; Pavan, M. "DRAGON", 2005.

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

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

[32] Pearlman, R.; Smith, K. Metric Validation and the Receptor-Relevant Subspace Concept. *J. Chem. Inf. Comput. Sci.* **1999,** *39,* 28–35.

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

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

[35] Broto, P.; Moreau, G.; Vandycke, C. Molecular Structures: Perception, Autocorrelation Descriptor and SAR Studies. *Eur. J. Med. Chem.* **1984,** *19,* 66–70.

[36] Moreau, G.; Broto, P. Autocorrelation of Molecular Structures: Application to SAR Studies. *Nouv. J. Chim.* **1980,** *4,* 757-764.

[37] Moreau, G.; Broto, P.; Fortin, M.; Turpin, C. Computer Conducted Screening of Molecular Structures of Potentially Anxiolytic Substances Using an Autocorrelation Technique. *Eur. J. Med. Chem.* **1988,** *23,* 275–281.

[38] Moran, P. Notes on Continous Stochastic Phenonema. *Biometrika* **1950,** *37,* 17–23.

[39] Geary, R. The Contiguity Ratio and Statistical Mapping. *Incorp. Statist.* **1954,** *5,* 115–145.

[40] Galvez, J.; Garcia, R.; Salabert, M.; Soler, R. Charge Indexes. New Topological Descriptors. *J. Chem. Inf. Comput. Sci.* **1994,** *34,* 520–525.

[41] Galvez, J.; Garcia-Domenech, R.; de Gregorio Alapont, C.; De Julian Ortiz, V.; Popa, L. Pharmacological Distribution Diagrams: A Tool for De Novo Drug Design. *J. Mol. Graphics* **1996,** *14,* 272–276.

[42] Galvez, J.; Garcia-Domenech, R.; De Julian Ortiz, V.; Soler, R. Topological Approach to Drug Design. *J. Chem. Inf. Comput. Sci.* **1995,** *35,* 272–284.

[43] Consonni, V.; Todeschini, R. *Rational Approaches to Drug Design;* Prous Science: Barcelona, 2001.

[44] 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.

[45] Consonni, V.; Todeschini, R.; Pavan, M.; Gramatica, P. Structure/Response Correlations and Similarity/Diversity Analysis By GETAWAY Descriptors. Part 2. Application of the Novel 3D Molecular Descriptors to QSAR/QSPR Studies. *J. Chem. Inf. Comput. Sci.* **2002,** *42,* 693–705.

[46] Schuur, J. H.; Selzer, P.; Gasteiger, J. The Coding of the Three-Dimensional Structure of Molecules By Molecular Transforms and Its Application to Structure-Spectra Correlations and Studies of Biological Activity. *J. Chem. Inf. Comput. Sci.* **1996,** *36,* 334–344.

[47] Gasteiger, J.; Sadowski, J.; Selzer, P.; Steinhauer, L.; Steinhauer, V. Chemical Information in 3D Space. *J. Chem. Inf. Comput. Sci.* **1996,** *36,* 1030–1037.

[48] Todeschini, R.; Lasagni, M.; Marengo, E. New Molecular Descriptors for 2D and 3D Structures - Theory. *J. Chemom.* **1994,** *8,* 263–273.

[49] Todeschini, R.; Gramatica, P. 3D Modelling and Prediction by WHIM Descriptors. Part 5. Theory Development and Chemical Meaning of WHIM Descriptors. *Quant. Struct.-Act. Relat.* **1997,** *16,* 113–119.

[50] Todeschini, R.; Gramatica, P. 3D-modelling and prediction by WHIM descriptors. Part 6. Applications of the WHIM descriptors in QSAR studies. *Quant. Struct.-Act. Relat.* **1997,** *16,* 120–125.

[51] Gramatica, P.; Corradi, M.; Consonni, V. Modelling and Prediction of Soil Sorption Coefficients of Non-Ionic Organic Pesticides by Different Sets of Molecular Descriptors. *Chemosphere* **2000,** *41,* 763–777.

[52] Todeschini, R.; Gramatica, P.; Marengo, E.; Provenzani, R. Modeling and Prediction by Using WHIM Descriptors in QSAR Studies: Submitochondrial Particles (SMP) as Toxicity Biosensors of Chlorophenols. *Chemosphere* **1995,** *33,* 71–79.

[53] Todeschini, R.; Vighi, M.; Provenzani, R.; Finzio, A.; Gramatica, P. Modeling and Prediction By Using WHIM Descriptors in QSAR Studies: Toxicity of Heterogeneous Chemicals on Daphnia Magna. *Chemosphere* **1996,** *32,* 1527–1545.

[54] Mattioni, B.; Jurs, P. Prediction of Dihydrofolate Reductase Inhibition and Selectivity Using Computational Neural Networks and Linear Discriminant Analysis. *J. Molec. Graph. Model.* **2003,** *21,* 391–419.

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

[56] Kier, L.; Hall, L. *Molecular Connectivity in Structure Activity Analysis;* Research Studies Press Ltd., John Wiley and Sons: Hertfordshire, England, 1986.

[57] Kier, L.; Hall, L. Molecular Connectivity I: Relationship to Local Anasthesia. *J. Pharm. Sci.* **1975,** *64,* 1971–1974.

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

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

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

[61] 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.

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