The SciTech, Journal of Science & Technology
Vol-1, Issue-1, p-3-39, 2012.
Article information:
Article Received: 24 March, 2012 Article Accepted: 25 March, 2012
Article Online: 16 May, 2012
Journal homepage: www.thescitechpub.com

Research Article

Article No: STJST201211001
DOI:……..

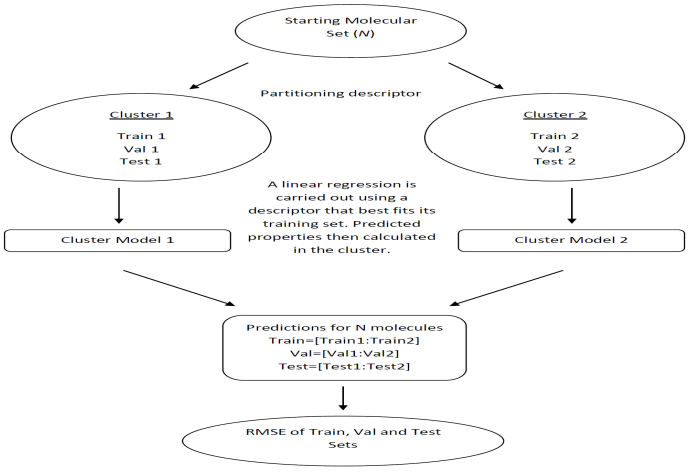
THE DEVELOPMENT OF MORE ACCURATE QSAR TECHNIQUES

A. Lee, A. G. Mercader, E. A. Castro*, P. R. Duchowicz
INIFTA (UNLP, CCT La Plata-CONICET), Diag. 113 y 64, Sucursal 4, C.C. 16, 1900 La Plata, Argentina
**Email: eacast@gmail.com, castro@quimica.unlp.edu.ar*

Abstract:

QSAR is a very effective starting step in the development of compounds for vast numbers of industries. Its scale and importance, especially in the medicinal field means it is a dynamic area to research. The size of QSAR also presents problems; there are many different methods in use for each of the steps in a study, from the descriptors in use, to the type of linear regression to apply to the descriptors. The idea was to put forward models that improved upon the existing methods to such a degree that it could become a universal method for QSAR modelling. This project successfully investigated in detail an improvement to the existing methods to choose the correct number of descriptors to include in the model by using R_{loo} analysis; this resulted in a simpler model compared to previous methods. K – Means clustering was also investigated as part of a novel, variable independent method. This methodology only uses one descriptor as opposed to general QSAR studies which use several. The results for 12 out of the 14 sets were at least as accurate as the results obtained by existing linear methods. An example using PERM; the S_{test} obtained using the novel method was 0.46 compared to the S_{test} of 0.53 obtained by using current linear methods. The simplicity associated with the K - Means clustering method and the fact it shows improved predictive potential could lead to an overhaul of all current, more complicated methods in favour of the simpler K- Means based method.

Graphical Abstract:



Content:

SL No	Topic
1	Introduction
1.2	Measuring the Accuracy of A QSAR Model
1.2.1	Standard Deviation
1.2.2	Correlation Coefficient
1.3	Datasets Used in This Work
2	The Representation of the Molecular Structure
2.1	Introduction
2.2	Graph Theory
2.2.1	Representing Graphs as Matrices
2.2.2	Connection Tables
2.2.3	Software Used to Represent the Molecular Structure
3	Molecular Descriptors
3.1	Introduction
3.2	0D Descriptors
3.3	1D Descriptors
3.4	1D Descriptors
3.4.1	2D Autocorrelation Descriptors
3.4.2	BCUT Descriptors
3.4.3	Galvez Topological Descriptors
3.4.4	Molecular Walk Counts
3.5	3D Descriptors
3.5.1	GETAWAY Descriptors
3.5.2	Randic Molecular Profiles
3.5.3	Morse Descriptors
3.5.4	Charge Descriptors
3.5.5	RDF Descriptors
3.5.6	WHIM Descriptors
3.6	Example of Variability of Molecular Descriptors
4	The Development of The QSAR Model
4.1	Introduction
4.2	Aims of search methodologies
4.3	Searching Methods
4.3.1	Full Search
4.3.2	Forward Stepwise Regression
4.3.3	Replacement Method (RM)
4.3.4	Modified Replacement Method (MRM)
4.3.5	Enhanced Replacement Method (ERM)
4.3.6	Comparison of RM, MRM and ERM
5	Validation
5.1	Introduction
5.2	External Validation
5.3	Internal Validation
5.3.1	Cross Validation Analysis
6	Determining the Optimum Number of Descriptors to Use in a Model
6.1	Introduction
6.2	Kubinyi FIT
6.3	VFIT
6.3.1	Development of VFIT Using R_{loo} Analysis of PKA set
6.3.2	VFITloo using RAD dataset
6.3.3	VFITloo using SINGLET OXYGEN dataset
6.4	Conclusion
7	K-Means Cluster Analysis
7.1	Introduction
7.2	K-Means Clustering
7.3	Partitioning the Dataset
7.4	Distance Measurement
7.4.1	City – Block Distance Measurement
7.4.2	Euclidean Distance Measurement
7.4.3	Comparison of Distances
7.4.4	Conclusion
7.5	The Proposal of A New Method of QSAR Modelling
7.5.1	Partitioning Method
7.5.2	Constitution of the Training, Validation and Test Sets
7.6	Method Description
7.6.1	Choosing the Value of *k*
7.6.1.1	Conclusion
7.6.2	Comparison with Existing Methods
7.7	Results
7.7.1	Greater than 60 Molecules
7.7.2	Fewer than 60 Molecules
7.8	Conclusion

1. Introduction:

A Quantitative Structure-Activity Relationship (QSAR) is the study of the dependence of the chemical structure on an observable experimental property or 'activity' over a collection of chemical compounds. Modelling this relationship allows predictions to be made about properties of previously unseen chemical compounds. Any QSAR study is based on the general equation- (1.1).

Activity=function (structural properties) (1.1)

One of the first historical applications of QSAR models was involved with the prediction of boiling points of straight chain alkanes[1]. It was noted that the boiling point increased with increasing length of the carbon chain, this allowed predictions to be made about higher length carbon chains without having to submit them for experimental analysis. Other notable QSAR works include the Hammett Equation and the related Taft Equation[2,3]. Recently both Biological properties, such as toxicity to organisms[4], and Physiochemical properties, such as aqueous solubility[5] have been modelled using QSAR analysis. Each of these studies is based on the QSAR fundamental assumption, that molecules with similar chemical structures exhibit similar activities[6]. This is countered by the Structure-Activity Relationship paradox that not all similarly structured molecules have similar activities. For the purpose of this QSAR project, the fundamental assumption will be taken as correct. Generally any attempt to model the relationship between an experimental activity and the chemical structure requires some type of regression analysis. This analysis requires a set of compounds with known activities to make the initial model. This is called the 'training set'. The regression extracts trends between the structures of these training compounds and the activity. The aim is to train an accurate enough model that permits the calculation of the activity of molecules with yet unseen.

Making the model requires each chemical structure to be represented in a form that can be easily analysed. This project will therefore be concerned with molecular descriptor based QSAR, where each molecule is represented by a set of chemical descriptors[7]; the origin and calculation of which will be explained in section 3. Each chemical structure is defined as the three dimensional disposition of atoms in a molecule, and by converting the structure to a matrix of descriptors each with fixed values, a mathematical connection can be formed between the activity values of the training set, and the values in the descriptor matrix.(Section 2)

Any QSAR model is limited by the variety of the training set molecules; therefore predictions on compounds with similar structure to the training set will be more reliable than predictions on molecules with marked differences from the training set. For example a training set of linear alkanes would not be accurate in predicting properties of a phenol series. This is known as the problem of the applicability domain in QSAR studies[8]. The applicability domain is the range of chemical structures that a QSAR model can be used to predict. Any structure with no relation to any structure in the training compound would be considered outside the domain of applicability of the model. The training set is where the regression algorithm finds genuine trends between the molecular structures and the experimental activities. However an overly enthusiastic regression on a training set will result in a situation known as overfitting[9] It is difficult to determine when a model will be overfitted, and it shall be investigated later in the project.(Sections 6 and 7)

The standard test for the model constructed from the training set is to use a test set of compounds whose activities are also known but have not been used in anyway in the calibration of the model. An accurate model will give similar predictions for the activities of the test set in relation to their genuine experimental activities[10].

An external test set is completely unrelated to the training set and consists of a set of compounds of random chemical structures each with the same measured activity parameter of the training set. An external set however may consist of molecules that are of completely different chemical structure to the training set so therefore are outside its domain of applicability and would not give an accurate picture of the predictive power of the model. The alternative is an internal test set, this consists of molecules that were originally in the complete set of starting molecules that were selected, and placed aside to validate the accuracy of the model. The molecules selected for the test set must be representative of the original complete set and should only be used after regression has been carried out on the training set. These have the advantage of having a higher probability of being in the applicability domain of the study.

Generally, an internal test set is smaller than the training set, acknowledging that despite the necessity of this validation step, it is merely present to evaluate the accuracy of the model produced by the larger training set. A large training set in comparison to the test set will allow a model to be produced using a higher diversity of molecules expanding the applicability domain.

The act of choosing the training and the test set is vital to accuracy and legitimacy of the QSAR model; it must be completely impartial, and any attempt to fix the test set to alter the results invalidates the model. In section 7

we will investigate the use of K-Means Cluster analysis to choose the molecular sets from the molecular pool for regression, and investigate whether this method of choosing the sets leads to more accurate results when compared with the pre-existing QSAR methods. Using K- Means analysis allows a large and varied dataset to be divided into smaller clusters which contain similar molecules, thus making it more flexible and easier to manage[10].

1.2 Measuring the accuracy of a QSAR model:

A QSAR model contains a variety of statistical measurements to determine the effectiveness of its predictive capability. The most important and most common are the standard deviation (*S*) and the correlation coefficient (R); how they relate to QSAR studies will be explained now.

1.2.1 Standard Deviation (*S*):

To demonstrate the meaning of *S* I have taken a selection of six compounds[11] with measured IC_{50} activity values for 1,1-Diphenyl-2-picrylhydrazyl (DPPH) free radical scavenging assay[12] This is the ability of each compound to reduce the concentration of the DPPH radical detected by 50% at 517 nm.

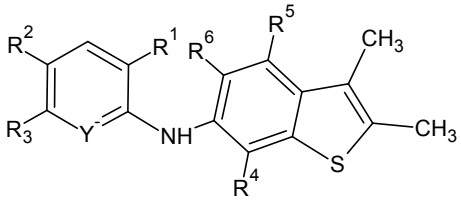


***Figure1.1**: Backbone of the molecules with measured DPPH*

***Table 1.1:** Different groups at different positions in the molecule affects the activity value*

Compo und. No.	R^1	R^2	R^3	R^4	R^5	R^6	Y	Experi-mental *-log(IC_{50})*
1	H	OCH_3	OCH_3	H	H	CH_3	C	4.3
2	H	OCH_3	H	CH_3	CH_3	H	C	3.89
3	H	OCH_3	H	H	H	CH_3	C	4.14
4	H	CHO	H	H	H	CH_3	C	2.39
5	Br	OCH_3	OCH_3	H	H	CH_3	C	3.81
6	Br	H	H	H	H	CH_3	C	3.05

With only the experimental values and the compound numbers, the best guess for the activity of any unknown compound is the mean of the activities of the known compounds, (Fig. 1.2) which in this case is 3.60

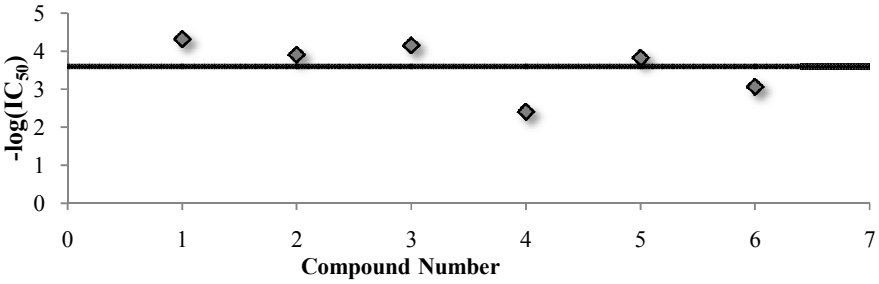


***Figure 1.2:** Activity values from table 1.1 and their mean*

S is defined as the dispersion of a set of data points from the mean of the data points; its equation for a sample of any number of compounds is,

$$S = \sqrt{\frac{\sum(x_i - \mu)^2}{N-1}} \tag{1.1}$$

where x_i is the activity of compound *i*, μ is the mean activity of the group of compounds and N is the total number of compounds. The *S* value for the group of six compounds in Fig. 1.2 is 0.732.

In QSAR the number of molecular descriptor in the model (*d*) is added to the *S* equation [13],

$$S = \sqrt{\frac{\sum(x_i - \mu)^2}{N-d-1}} \tag{1.2}$$

The molecular descriptor chosen for this example is a Harary H Index descriptor [14]. This is defined as the half-sum of the off diagonal elements of the reciprocal distance matrix $[D^r]_{ij}$.(Section 2.2.1)

$$H = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}[D^r]_{ij} \tag{1.3}$$

where the reciprocal distance matrix is derived from the standard distance matrix as follows,

$$[D^r]_{ij} = \frac{1}{[D]_{ij}} \tag{1.4}$$

The *H* values for this molecular set are shown in Table 1.2. and are also shown as a plot of DPPH assay vs. Harary Index in Fig 1.3.

***Table 1.2**: Compound numbers, their corresponding activities and Harary Index*

Compound Number	DPPH assay *(-log(IC_{50}))*	Harary Index (Har)
1	4.3	42.941
2	3.89	41.237
3	4.14	38.731
4	2.39	38.371
5	3.81	45.329
6	3.05	39.648

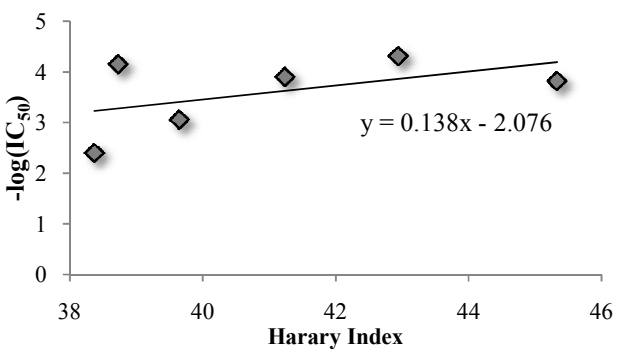


***Figure 1.3**: Experimental –log(IC_{50}) vs Harary Indices*

A linear regression gives a line of best fit through the points obtained by plotting Har vs. Activity, this is the best predictive model for the data available and can give a predicted IC_{50} value for any *Har* value. The obtained equations is the following,

$$-\log\,(IC_{50})_{pred} = 0.138Har - 2.07 \tag{1.5}$$

Table 1.3 *:Experimental and predicted activity values; and the difference (residuals) between them*

Experimental – log(IC_{50}) values	Predicted – log(IC_{50}) values	Residuals
4.3	3.850	0.450
3.89	3.615	0.275
4.14	3.269	0.871
2.39	3.219	-0.829
3.81	4.179	-0.369
3.05	3.395	-0.345

When this values are used on equation (1.2) *S* equals 0.630. This is smaller and therefore shows a more effective linear regression between the 6 molecules than using barely the mean of the experimental values. Thus showing that by correlating certain molecular descriptors with the experimental values, a more precise relationship can be made between the activities of the compounds and their molecular structure. Now to predict a DPPH assay activity for any unknown compound using this model, only the *Har* value is needed, which can easily be determined. These predicted results can be plotted against the experimental results to give a graph showing the accuracy of the predictions; where complete correlation would lay on the line of best fit indicating the same predicted and the experimental values.

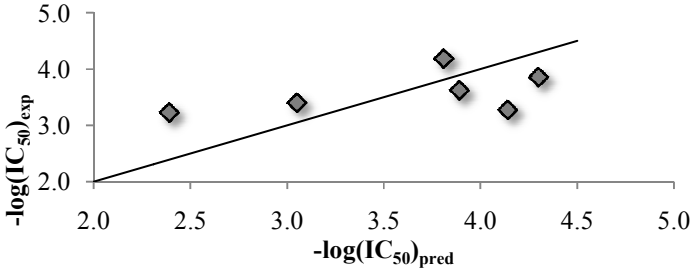


Figure 1.4*: Predicted vs. experimental values, line of y = x indicates perfect correlation*

1.2.2 Correlation Coefficient:

As well as the *S*, another very important statistic parameter in QSAR model analysis is the correlation coefficient (R)[15]. R is a measure of how well all of the data points are correlated to the line of best fit of the model. R values always fall between 0 and 1, where 1 demonstrates complete correlation with the line of best fit and 0 indicates no correlation at all. It has the formula,

$$R^2 = \frac{Regression\ Variance}{Original\ Variance} \tag{1.6}$$

where the original variance is the sum of squares of deviations from the mean, which for this dataset is 2.68. (Fig. 1.3)

$$\sum(x_i - \mu)^2 \tag{1.7}$$

The regression variance is the sum of squares of deviations from the mean, subtract the sum of squares of residuals. For this dataset this sum of squares of residuals is 1.98. (Table 1.3) This means the R^2 of this dataset is 0.7 divided by 2.68 which equals 0.261. This compares favourably with the R^2 of the mean model which is 0.183.

This is an example using just one molecular descriptor chosen at random, most QSAR models use more than one descriptor selected from over 1500 available, leading to very accurate model for full QSAR studies[16,17].

Figure 1.4 can be used as the base for the model validation step by using the calculated values for a test set. Placing the calculated points on Fig. 1.4 and calculating *S* gives the accuracy of the test set with respect to the training of the model.

The role of the molecular descriptors is vital in every facet of QSAR, and t will be explored in this text. A thorough description and evaluation of these molecular descriptors and how they represent the molecular structure is required before any further developments.

1.3 Datasets used in this work

- 166 Molecules with measured aqueous solubility (SOL)[18].
- 470 Molecules with measured growth inhibition of the ciliated protozoan *Tetrahymena pyriformis* (TOX)[19].
- 128 Molecules with measured inhibition of HIV-1 reverse transcriptase (HIV)[20].
- 41 Molecules with measured deactivation of singlet oxygen rate constant (SINGLET OXYGEN)[21].
- 52 molecules with measured DPPH radical scavenging assay (RAD)[11,22].
- 116 molecules with measured fluorescence values (FLUOR)[23].
- 100 flavonone derivatives with measured ED_{50} values (MES)[24,25].
- 75 Thiadiazoloidine derivatives with measured inhibition of serine proteases (THIA)[26].
- 80 pharmaceutical compounds with measured dissociation constants (PKA)[27].
- 78 flavonanone derivatives with measures GABA inhinbition (GABA)[28].
- 70 drugs with measured pharmacological permeabilities (PERM)[29].
- 22 compounds with measured anti - malarial activities (MALAR)[30].
- 41 naftoquinones with measured anti cancer activity (NAFT)[31].

2 The representation of the molecular structure:

2.1 Introduction:

In QSAR the molecular structure is defined as the array of the atoms and bonds that constitute a molecule[32]; and it is responsible for many of the measurable properties of the molecule in question. It can be described in many different ways; from nomenclature, to diagrams showing the 3D structure of the molecule including the outer electron clouds. The standard representation of molecules is a 2D representation where atoms are represented by their atomic symbol and the bonds are represented by lines. This is however only an incomplete and highly simplified visual aid; it merely shows the topology of the molecule with no reference to the 3D topography. The three dimensional structure requires additional information such as the coordinates of atoms in space, the bond angles and the bond distances. All these different aspects of the structure of molecules can be translated to a numerical code, known as molecular descriptors[7]. There are thousands of available molecular descriptors available for QSAR studies; however this project will deal with the descriptors from the “Dragon” descriptor program[33] as well as descriptors obtained during the structure optimization step using HyperChem[34]. Many of the molecular descriptors are developed from graph theory.

2.2 Graph Theory:

Graph theory is used in mathematics to describe a great variety of problems and situations, [35] the first and most famous example being Euler’s paper on The Bridges of Konigsberg in 1736[36]. The connection between graph theory and chemical structure is the basis for developing many different sets of descriptors for use in QSAR analysis. In mathematical terms, chemical diagrams can be considered as common graphs. These graphs consist of vertices that represent the atoms of the molecules and edges that represent the bonds in the molecule. This type of graph is defined as a topological graph and it merely shows the connection between atoms and the type of bond, it contains no geometric information.

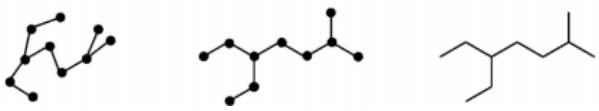


Figure 2.1: *Three representations of the same structure according to graph theory*

A weighted graph contains numbers or symbols on its vertices unlike in Fig. 2.1. They can also have more than one edge connecting each node; which in chemical graphs represent multiple bonds.

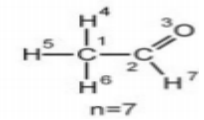


Figure 2.2: *Acetaldehyde molecular graph weighted with atomic symbols.*

2.2.1 Representing Graphs as Matrices:

A graph obtained using graph theory can be represented in number matrix form. The advantage of this is that matrix calculations are very well understood and easily carried out. This finally gives a direct link between chemical structure and numerical reasoning. The matrix to be shown is the bond adjacency matrix for acetaldehyde. This is a matrix formed using 1's and 0's (bits), where 1 signifies that a bond exists between the two atoms it represents in the matrix, and 0 signifies no bond.

a	**1**	**2**	**3**	**4**	**5**	**6**	**7**
1	*0*	*1*	*0*	*1*	*1*	*1*	*0*
2	*1*	*0*	*1*	*0*	*0*	*0*	*1*
3	*0*	*1*	*0*	*0*	*0*	*0*	*0*
4	*1*	*0*	*0*	*0*	*0*	*0*	*0*
5	*1*	*0*	*0*	*0*	*0*	*0*	*0*
6	*1*	*0*	*0*	*0*	*0*	*0*	*0*
7	*0*	*1*	*0*	*0*	*0*	*0*	*0*

Figure 2.3(a): *Connectivity matrix of acetaldehyde corresponding to Fig. 2.2*

b	1	2	3	4	5	6	7
1		*1*		*1*	*1*	*1*	
2	*1*		*1*				*1*
3		*1*					
4	*1*						
5	*1*						
6	*1*						
7		*1*					

Figure:2.3(b) *Connectivity matrix of acetaldehyde omitting ceros*

These two matrices are identical; *a* shows the bonds between the atoms as well as describing where there is no bond, *b* is a simplified matrix omitting all 0's. An even simpler matrix is shown in Fig 2.4 which omits all hydrogen atoms leaving only carbons and oxygen; and also leaves out duplicated information.

	1	**2**	**3**
1		1	
2			1
3			

Figure 2.4: *Connectivity matrix of acetaldehyde, omitting all H atom*

There are many different types of matrix that can be used to represent the molecular structure, they have both advantages and disadvantages. A summary was included in Table 2.1.

Table 2.1: *Advantages and disadvantages of the different type of matrices that represent the molecular structure*

Advantages	Disadvantages
General Matrix: General aspects of the representation matrix	
• It completely codifies the molecular diagram (each atom and bond is represented) • Can be subjected to matrix algebra	• The number of entries in the matrix grows with the square of the number of atoms • Stereochemistry is not represented
Adjacency Matrix: Rows and columns represent the atoms (vertices)	
• Describes the connections of all the atoms • Only contains zeros and ones (bits)	• Doesn't represent the type or order of the bonds • Doesn't represent the number of free electrons
Distance Matrix: Distance is expressed as a geometric (Å) or topological (number of bonds between atoms)	
• Describes the geometrical distance • Easily changed to 3D distance matrix by using actual distances between atoms rather than number of bonds	• Doesn't represent the type nor the order of the bonds • Doesn't represent the number of free electrons • It is not represented by bits
Incidence Matrix: The columns represent the atoms (vertices) and the rows represent the bonds (edges)	
• Describes the connections and bonds • Only contains zeros and ones (bits)	• Doesn't represent the type nor the order of the bonds • Doesn't represent the number of free electrons
Bond Matrix: The columns and rows represent the atoms (vertices). Any multiple bond is represented by the corresponding number, e.g. Double bond is represented by a 2	
• Describes the connections and order of the bonds	• Doesn't represent the number of free electrons
Bond – Electron Matrix: Is considered as an extension to the bond matrix. It numbers the all the valence electrons in the molecule.[37]	
• Describes the connections and orders of the bonds and the number of valence electrons	• Not represented by bits

2.2.2 Connection Tables:

These are an alternative form of representing the molecular structure to matrices, it consists of a table where the atoms occupy the rows and each column presents information about those bonds of each atom. These tables have been used as the main way of representing molecules in computer systems; they can be applied equally as well as graph theory matrices in the development of molecular descriptors.

Table 2.2: *Connection Table of Acetaldehyde*

Atom Number	Element	Connected with:	Bond order	Connected with:	Bond order	Connected with:	Bond order	Connected with:	Bond Order
1	C	1	1	4	1	5	1	6	2
2	C	1	1	3	2	7	1		
3	O	2	2						
4	H	1	1						
5	H	1	1						
6	H	1	1						
7	H	2	1						

2.2.3 Software used for the representation of the molecular structure:

Before calculating any 3D molecular descriptors, each molecular structure must be optimized. This is adjusting the 2D structure of the molecular into a more energetically favourable 3D structure, representing much more accurately the natural structure of each compound. Also this allows 3D descriptors to be calculated direct from the structure. This thesis used the Hyperchem program[34] to first input the 2D structures of the molecules, then to optimize the structure to a lower energy conformation. There are two steps used for the optimization in this work, first a pre - optimization is carried out using *Molecular Mechanics Force Field* (MM+), then a refinement step is carried out using the semi - empirical *Parametric Method 3* (PM3) which uses the Polak–Ribiere algorithm[38].

Table 2.3: *Connection table obtained for acetaldehyde after Hyperchem optimization*

Atom N°	Element	Atom Type	Atomic Charge	Cartesian Coordiantes			N° of bonds	Bonded atoms and type of bond							
				x	y	z									
1	C	sp^3 C	-0.1959	-0.2601	-1.0482	2.080E-06	4	2	S	4	S	5	S	6	S
2	C	sp^2 C	0.2811	-0.2882	0.4522	-3.690E-06	4	1	S	3	D	7	S		
3	O	O	-0.3182	-1.3141	1.0939	-2.580E-07	2	2	D						
4	H	H	0.0526	0.7675	-1.4317	-2.720E-06	1	1	S						
5	H	H	0.0695	-0.7677	-1.4579	8.836E-01	1	1	S						
6	H	H	0.0695	-0.7677	-1.4579	-8.836E-01	1	1	S						
7	H	H	0.0415	0.6821	0.9747	2.130E-07	1	2	S						

3 Molecular Descriptors:

3.1 Introduction:

An indispensable step in any QSAR study is the calculation of molecular descriptors; this is the final stage in representing the molecular structure as a numerical code that can be used for
QSAR studies. The descriptors are classified into different families depending on both their dimensionality and what they represent in each molecule.

3.2 0D descriptors:

0D descriptors describe the constitution of the molecule and are independent of the connectivity and molecular conformation; the most obvious are the number of atoms in the molecule, bond types, molecular weight and average atomic weight. The utility of these descriptors can be appreciated when examining the number of hydrogen atoms in a molecule for example. A molecule with a large number of hydrogens in it would have a higher possibility of hydrogen bonding, whose effects on the activities of biological systems are notorious[39].

3.3 1D descriptors:

These are descriptors that involve fragments of the molecule and subsets of atoms. Atomic subsets include the number of sp^3 carbon atoms, number of cyanide groups, etc. Molecular fragments are counts of distinct fragments of each molecule; examples include hydrogens bonded to heteroatoms, and fluorine atoms bonded to sp^3 carbon atoms as well as Ar–OH fragments, etc. Functional group descriptors are a member of this dimension of descriptors, they are generally very important in QSAR works as many of the properties of molecules and their reactions occur due to the presence of functional groups[39].

3.4 2D descriptors:

These are molecular descriptors obtained from graph theory (section 2.2.1), independent of the molecular conformation.

3.4.1 2D Autocorrelation Descriptors:

2D autocorrelations are bi-dimensional correlations between pairs of atoms in a molecule, and they were defined in order to reflect the contribution of a certain atomic property to the experimental property under observation[40]. Many different atomic properties can be used to differentiate between the atoms in the molecule including mass, polarisablity, electronegativity or the atomic volume. The distances between the two atoms under study are calculated from the distance matrix, where the distance is the number of bonds between the two atoms selected. These distances are then weighted using one of the atomic variables mentioned,

$$\mathrm{ATS}dw = \sum_{i=1}^{n}\sum_{j=1}^{n}\delta_{ij}(w_i w_j) \qquad (2.1)$$

where w is the atomic weighting property, and δ_{ij} is the delta ratio which is $\delta_{ij} = 1$ if $\delta_{ij} = d$ and $\delta_{ij} = 0$ if δ_{ij} is not d (bond distance).

3.4.2 BCUT Descriptors:

These descriptors are values obtained from a modified connectivity matrix called the Burden Matrix[41]. In a molecular graph with the hydrogen atoms removed is defined as follows:

- The diagonal entries of the matrix are the atomic numbers of the elements.
- Off diagonal entries representing bonded atoms are equal to $\pi^*.10^{-1}$ where π^* is the bond order (i.e. 1, 1.5, 2 for single, aromatic and double bonds respectively)
- An additional 0.01 is added to off diagonal entries corresponding to terminal bonds.
- Any remaining entries are fixed to 0.001

To obtain different BCUT descriptor, each matrix is weighted using a property of each atom in the matrix, for example atomic mass, electronegativity, and polarisability.

3.4.3 Galvez Topological Charge Indices:

These indices describe the transfer of charge between pairs of atoms, and therefore of the global transfer of charge in the molecule. The Galvez Matrix ($\boldsymbol{M}$) is defined as,

$$\boldsymbol{M} = \boldsymbol{A}.\boldsymbol{D}^{-2} \qquad (2.2)$$

where $\boldsymbol{A}$ is the adjacency matrix and and $\boldsymbol{D}^{-2}$ is the reciprocal square of the distance matrix. $\boldsymbol{M}$ is a charge transfer matrix of a x a where a is the number of atoms in the molecule[42,43]. The charge transfer CT_{ij} is defined as δ_i for the matrix when $i = j$, and m_{ij} - m_{ji} when i and j are different. The diagonal elements of $\boldsymbol{M}$ are the vertex degrees of the atom (the number of atoms next to it in the H - depleted molecular graph), and the off diagonal values are size of the charge transfer from atom i. These descriptors are very effective at showing the distribution of charge in a molecule so are very good at modelling dipole moments of aromatic and unsaturated hydrocarbons[44].

3.4.4 Molecular Walk Counts:

These are descriptors obtained by counting walks of different lengths both outward and returning walks. For example in methanol there are is a maximum length of molecular walks of 3 bonds, or a return molecular walk of 6 bond lengths. There are many other molecular walks of 1 bond length and 2 bond lengths.

3.5 3D descriptors:

3D molecular descriptors take into account the conformation of the molecular structure, bond distances, bond angles, dihedral angles, etc. They are therefore able to describe the stereochemical properties of the molecules. The molecular matrix (***M***) is formed of the Cartesian Coordinates of each atom in the optimized molecular structure, which is calculated from the geometric centre of the molecules[44]. The 3D descriptors are generally calculated from the molecular matrix or the distance matrix (Table 2.1). A problem with 3D descriptors that is not present in other dimensional descriptors is that they are obtained from the optimized conformation of each molecule. The optimized conformers are not the lowest possible energy conformers of the molecules due to quantum mechanical approximations made in both the semi empirical PM3 method[45] and the MM+ molecular mechanics [46] method when optimizing the structure.

The GETAWAY (GEometry Topology and Atom-Weights assemblY) descriptors try to match 3D molecular geometry provided by the 3D molecular influence matrix and atom relatedness from molecular topology, with chemical information by using different atomic weightings, such as atomic mass, polarisability, electronegativity and Van der Waal's volume[47] The molecular influence matrix (***H***) is defined as,

$$H = M.(M^T.M).M^T \tag{2.3}$$

where ***M*** is the 3D Cartesian molecular matrix and $\boldsymbol{M^T}$ is the transpose of the molecular matrix. The diagonal elements of matrix ***H*** (h_{ij}) represent the influence of each atom on the molecular shape. For example atoms towards the exterior of the molecule have larger h_{ij} values than molecules closer to the centre of the diagonal. This supports the general idea that atoms on the outer edge of the molecule have more influence on its manifestation than central atoms. The off diagonal h_{ij} values also relate the accessibility of atom *j* to the atom *i*.

3.5.2 Randic Molecular Profiles:

Randic Molecular Profiles are molecular descriptors derived from the distribution of the distance matrix, defined as the average row sum of the matrix entries. Different descriptors are then obtained by raising the average row sum to the power *k* then normalising the results using *k!*. Since they are calculated using the 3D distance matrix they are a numerical representation of the entire structure of the molecule[48].

3.5.3 3D Morse Descriptors:

These descriptors are a representation of the 3D structure of the molecule in question based on electronic diffraction, providing 3D information of the structure using a transformation based on the electronic diffraction equation. (2.4) Various properties of the molecule can be taken into account given the high flexibility of this representation of the molecule[49].

$$I(s) = \sum_{i=2}^{A} \sum_{j=1}^{i-1} w_i . w_j \frac{sen(s.r_{ij})}{s.r_{ij}} \tag{2.4}$$

where *I*(*s*) is the intensity of the electronic dispersion from a reciprocal distance *s*, *A* is the number of atoms in the molecule, *w* is the atomic property used to weight the result. This allows numerical representation of the three dimensional distribution of distinct atomic properties in a molecule. r_{ij} is the inter atomic distance between atoms *i* and *j*.

3.5.4 Charge Descriptors:

Charge descriptors describe the distribution of charge in the molecule and are only calculated after the optimization step to ensure closest possible correlation to experimental values. Examples are the sum of atomic charges, total squared charged, maximum positive charge, HOMO and LUMO energies ect. The HOMO descriptor is the energy of the highest occupied molecular orbital, while the LUMO descriptor corresponds to the energy of the lowest unoccupied molecular orbital. Each descriptor describes the reactivity of the molecule; the HOMO describes the susceptibility of attack from electrophiles and the LUMO describes the susceptibility of attack from nucleophiles. Lewis bases donate electrons from their HOMO, so the strength of a Lewis base increases with the

increase in HOMO energy, whereas the strength of a Lewis Acid decreases with an increase in energy of the LUMO. A large energy gap between the HOMO and LUMO of the molecule is associated with a very high stability.

3.5.5 RDF Descriptors:

Radial Distribution functions describe how atomic density varies as a function of distance from one particular atom. It is defined as the probability of finding an atom in spherical radius *r,* where generally the distance *r* is varied ±1 between 0.5Å and 15.5Å starting at a specified atom. Also incorporated are different atomic properties in order to differentiate between the contributions of each atom to the property under study. For example the descriptor RDF010*e* is the atoms found in the a radius of 10Å, with each atom weighted with its electronegativity to distinguish its nature[50].

3.5.6 WHIM Descriptors:

Weighted Holistic Invariant Molecular (WHIM) Descriptors are 3D molecular indices that represent different sources of chemical formation[50]. They are calculated from the projection of atoms along the principal molecular axis and contain information about the entire 3D structure, such as shape, total volume, and symmetry. Weighted connection tables containing the *x, y* and *z* coordinates (table 2.3) are used to calculate the numerical values.

$$s_{ij} = \frac{\sum_{i=1}^{n} w_i(q_{ij}-Þ)(q_{ik}-Þ)}{\sum_{i=1}^{n} w_i} \tag{2.6}$$

where n is the number of atom, w_i is the weighting property of the ith atom, q_{ij} and q_{ik} are the coordinates of the jth and k^{th} atoms in the molecule. Þ is the average distance from atom i from atom j. S_{ij} is a 3 x 3 matrix where the elements are the weighted covariances between the atoms j and k[51,52].

3.6 Example of variability of molecular descriptors:

Table 3.1: *S and R of QSAR models using different descriptor family (using RAD set) For each family only 1 descriptor was used to train the model.*

Descriptor family	**S**	**R**	**Descriptor family**	**S**	**R**
BCUT	1.30	0.18	Randic Molecular Profiles	1.28	0.25
2D Autocorrelations	1.20	0.40	Galvez Charge Transfer	1.25	0.33
Molecular Walk Indices	1.31	0.14	Functional Group Indices	1.23	0.36
Topological Descriptors	1.23	0.37	Atom Centred Descriptors	1.11	0.54
GETAWAY	1.16	0.48	Charge Descriptors	1.28	0.24
Morse Descriptors	1.22	0.39	Constitutional Descriptors	1.05	0.61
WHIM	1.39	0.19			

According to Table 3.1 the most effective descriptor family for training the 52 radical scavengers is the 0D constitutional descriptor family. The descriptor used to train the molecule with this accuracy was *Me*; which is the mean electronegativity of the molecule, where the electronegativity of C is fixed at 0. This could mean that the electronegativity of atoms in the molecule is relevant in the molecule's radical scavenging activity. This is a sensible conclusion because molecules with high average electronegativity will contain heteroatoms, these can provide resonance stabilisation of the radical compound formed in the DPPH radical scavenging assay. A larger average

electronegativity would generally indicate a greater ability to provide resonance stabilisation to the radical compound, the more resonance forms available the more stable the radical compound and therefore the more likely it is to accept a radical electron[52].

4 The development of the QSAR model:

4.1 Introduction:

The development of the model involves two fundamental steps; selecting which descriptors to use, and the optimum number that would give the most accurate model. The mathematical function used to search should also be selected, be it linear or non linear. There are generally three types of model development techniques; type 1 involves using Multiple Linear Regression (MLR) for both steps, making it entirely linear, type 2 uses MLR to chose the descriptors but a non linear method such as artificial neural networks[53] to carry out the calibration. Type 3 uses non linear methods for both the selection and the calibration steps.

4.2 Aims of search methodologies:

Each method finds an optimum number of descriptors (d) from a pool that contains the total number of descriptors (D) where $D>>d$, in the form $\mathbf{d} = (d_1.....d_f)$. Using these descriptors the values of S and the R of the model can be obtained.

4.3 Searching Methods:

In this section a detailed description of the methods used in this work and how they differ in the accuracies of their results is presented. Only linear methods have been used in this text, and these have been optimised in order to give the most accurate results possible within the boundaries of acceptable computing time. These results compare similarly with non linear methods and are more informative[54]. Non linear methods such as Artificial Neural Networks process the information from the molecular structures and give out results; they do not show any relation between the molecular structure and the results obtained[53]. This is called the 'black box' problem[55]. Linear methods allow this relationship to be easily analysed and gives information about the model obtained, such as equation (4.2). An example of linear regression was introduced in the introduction, albeit with just one descriptor variable.(1.5) MLR uses more than one variable to connect the activity with the descriptors which allows more accurate models to be obtained.

4.3.1 Full search:

The full search method tries all available cases in order to deduce the best model of d descriptors from D total descriptors, therefore all posible regressions are carried out. This gives a total of (4.1) linear regressions.

$$\frac{D!}{d!((D-d)!)} \tag{4.1}$$

Due to the huge amount of regressions and time needed during the full search with a typical value of D, it is necessary to apply approximations and simplifications to the full search.

4.3.2 Forward stepwise regression:

This is a method derived in the 1950's and has been used in many QSAR studies to develop the model[56]. A forward regression starts off with no variables in the model. The desired number of descriptors is inputted manually, and the FSR tries descriptors from the descriptor pool, in each step including one at a time in the model. The regression then ceases when the number of steps taken equals the number of d. For example starting from an empty model the regression analysis algorithm adds a descriptor from the available total set D, which gives the lowest S. It then moves on and adds a second descriptor from the remaining D-1 set to the formula, which gives the lowest S again. This carries on until the amount of descriptors added is equal to the desired amount of descriptors inputted in the algorithm. This method does not give the lowest possible S of the model because only one variable is changed at a time. This is to say adding certain descriptors may decrease the overall S but may increase the standard

error of some of the descriptors determined earlier in the model; descriptors which could be altered again and further increase model accuracy. For this reason the following algorithms were developed to allow more flexible replacement of descriptors.

4.3.3 Replacement Method:

The replacement method is a multi variable linear regression method that allows the user to analyse hundreds of descriptors at once in order to provide the best selection of variables for a model. Its main advantage is that it gives similarly accurate results to the full search method without using the vast amount of linear regressions and computational time[57]. The idea of the replacement method is to take into account the relative *S* of the regression coefficients in order choose the descriptors and thus to achieve a lower *S*. Using the errors of these coefficients constitutes a way of choosing the descriptors to be replaced in the linear equation[57].In a descriptor model of *d* starting descriptors from *D* total descriptors would consist of,

$$S = \{\alpha d_1 + \beta d_2 + \gamma d_3 \ldots + \omega d_d\} \tag{4.2}$$

The replacement method algorithm starts with the manual input of a starting set of descriptors, for example [d_1 d_2 d_d]. The regression coefficients are determined for each of the *d* descriptors and then the algorithm first replaces d_1. This descriptor is replaced with each one of the remaining *D-d* descriptors. A linear regression is carried out after each replacement and the descriptor that corresponds to the lowest *S* value is kept. This is called 'one step' and the new equation is kept.
Next the descriptor with the highest coefficient value is replaced with every remaining descriptor, omitting the one previously replaced. Again a linear regression is carried out after each replacement and the descriptor that gives the lowest *S* is kept. This step is applied in turn to each descriptor with the largest coefficient value until each descriptor has been replaced once, this is called one round. The method carries on until replacing all of the d_d descriptors in a round doesn't result in a lower *S*, it then finishes, giving the lowest *S* possible from this method. The method can then be applied starting at another *d* rather than d_1. This may give different final *S* values depending on the starting descriptor.

The RM is an iterative method which gives very good *S* values in very short calculation times. Its accuracy can at times be limited when it falls into a local minimum of *S* from which it cannot break out. There are therefore improvements than can be incorporated into this algorithm to make it more accurate.

4.3.4 MRM–Modified Replacement Method:

This method is very similar to the RM but aims to reduce the probability of the *S* of the QSAR model getting caught in a local minimum, thus giving a more accurate model. In the RM model, the finishing point occurs when changing each descriptor in turn doesn't produce a lower *S*. In the MRM the algorithm replaces the descriptor d_x with the highest regression coefficient, with the best descriptor d_y even if it temporally gives a higher value of *S*. Afterwards, as the algorithm continues, the momentarily rise in *S* is overcome and an even lower value of *S* is obtained. If the algorithm does not converge after 350 steps have been carried out, it is forced to stop[58]. Below shows the evolution of the MRM method as a graph,

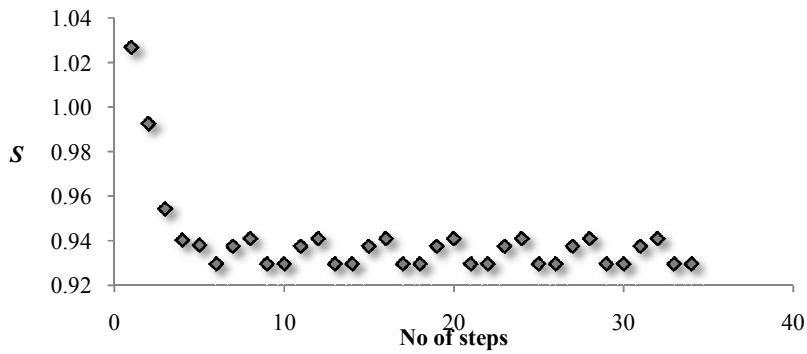


Figure 4.1: *Development of the MRM*

4.3.5 ERM – Enhanced Replacement Method:

RM and MRM have different finishing points that are independent of each other. This allows us to use different combinations of the RM and MRM together to try and achieve an even lower *S* than the MRM or RM alone. Table 4.1 shows us that a lower *S* can be obtained by combining the different search methods in the form RM – MRM – RM, known as the ERM[58].

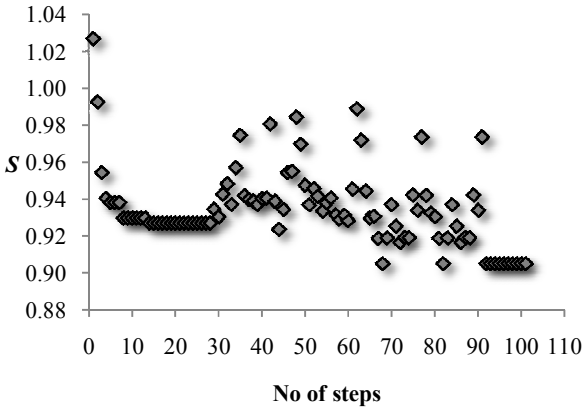


Figure 4.2: *Development of the ERM*

Table 4.1: *Descriptor selection, S and number of steps of the RM, MRM and ERM*

Step	d_1	d_2	d_3	d_4	d_5	d_6	d_7	S
1	1	2	3	4	5	6	7	1.0267521
2	70	2	3	4	5	6	7	0.99243693
3	70	2	3	4	5	6	40	0.95421562
4	70	61	3	4	5	6	40	0.94018769
5	70	61	50	4	5	6	40	0.93795998
6	70	61	50	4	5	6	40	0.93795998
7	70	61	50	4	5	6	40	0.93795998
8	70	78	50	4	5	6	40	0.92953888
9	70	78	50	4	5	6	40	0.92953888
10	70	78	50	4	5	6	40	0.92953888
11	70	78	50	4	5	6	40	0.92953888
12	70	78	50	4	5	6	40	0.92953888
13	70	78	50	4	5	6	40	0.92953888
14	70	78	50	11	5	6	40	0.92682368
....								
28	70	78	50	11	5	6	40	0.92682368
End of the RM (1), now the MRM element starts.								
29	70	78	50	11	5	6	18	0.93433662
30	70	61	50	11	5	6	18	0.93022416
31	70	61	3	11	5	6	18	0.94248581
32	70	61	3	11	28	6	18	0.94813444
33	70	61	3	11	28	4	18	0.93690899
34	70	61	3	13	28	4	18	0.9568875
35	76	61	3	13	28	4	18	0.97443479
36	76	47	3	13	28	4	18	0.94166135
37	76	47	3	13	40	4	18	0.93960852
38	76	47	3	13	40	4	41	0.93888404
39	76	47	3	13	40	8	41	0.93690848

40	71	47	3	13	40	8	41	0.94023656
...								
67	40	5	69	13	45	74	80	0.91856599
68	40	5	69	13	45	30	80	0.90483531
69	40	5	69	13	45	30	72	0.9188276
70	40	5	69	11	45	30	72	0.93685907
71	40	5	69	11	45	52	72	0.92515569
72	40	5	69	11	45	52	80	0.91641636
73	40	5	69	11	67	52	80	0.91907025
74	45	5	69	11	67	52	80	0.91891157
75	45	5	74	11	67	52	80	0.94209109
76	45	5	74	13	67	52	80	0.93380056
77	45	3	74	13	67	52	80	0.97343099
78	5	3	74	13	67	52	80	0.94192329
79	5	40	74	13	67	52	80	0.93234433
80	5	40	74	13	45	52	80	0.93030379
81	5	40	74	13	45	69	80	0.91856599
82	5	40	30	13	45	69	80	0.90483531
The MRM has decreased the standard deviation to a lower level than RM but is now repeating, (2) the RM starts again.								
....								
111	5	40	30	13	45	69	80	0.90483531
112	5	40	30	13	45	69	80	0.90483531
The RM cannot decrease the standard deviation any lower, the algorithm now terminates. (3)								

In Fig. 4.2 it is clear that the end points of each of the RM – MRM – RM algorithms, at points 1, 2 and 3 respectively. The primary RM algorithm gives a final *S* of 0.926, the MRM takes over then to try and overcome the local minimum of *S* the RM is in. The points in the MRM part of the graph show that when the descriptors corresponding to the low model *S* are replaced with descriptors that increase the *S*, eventually a set of new descriptors can be found that decreases the *S* lower than the RM value; in search of a new combination of descriptors that gives a lower *S* than was obtained with the RM method. This is achieved at step 68. The RM is then reapplied to the set of descriptors that achieved this lower value of *S* but it does not decrease the *S* anymore.

4.3.6 Comparison between the three methods RM, MRM and ERM:

The accuracy of the RM MRM and ERM algorithms was tested to show the aforementioned explanations. Each different method was tried with 2 different datasets,

Table 4.2: *A comparison of the accuracy of each method and calculation time*

RAD					
d	**1**	**2**	**3**	**4**	**5**
RM (S)	0.5012	0.4393	0.4064	0.3742	0.3570
MRM (S)	0.5012	0.4168	0.3720	0.369	0.3660
ERM (S)	0.5012	0.4168	0.3830	0.367	0.3510
Full Search (S)	0.5012	0.4168	0.3723	0.3617	0.3158
Time (RM) /s	0.0625	0.2656	0.4688	0.7969	1.2813
Time (MRM) /s	0.0313	0.3438	0.7344	25.4063	1.9219
Time (ERM) /s	0.0313	1.2969	1.5313	3.2500	6.7188
Time (FS) /s	0.0313	0.4688	16.5156	285.7813	5.04E+03
SOL					
d	1	2	3	4	5
RM (S)	1.2641	1.0099	0.9935	0.9693	0.9358

MRM (S)	1.3756	1.0099	0.9806	0.9604	0.9445
ERM (S)	1.2641	1.0099	0.9806	0.9455	0.9281
Full Search (S)	1.2641	1.0099	0.9806	0.9455	0.9206
Time (RM) /s	0.0416	0.5156	1.2344	2.3750	3.8954
Time (MRM) /s	0.0156	0.5000	1.0156	1.9844	25.0516
Time (ERM) /s	0.0625	1.7344	4.9219	7.8906	26.7969
Time (FS) /s	0.0547	0.8438	25.5315	458.0781	7.69E+03

The general trend in both datasets is to have an increase in accuracy from RM < MRM < ERM < FS, this however comes with a time penalty. The FS time increases to such an extent that it becomes impractical for larger descriptor sets and with more descriptors in the model. The ERM model is the most accurate of the remaining models and does have an extremely large time penalty. The RM, MRM and ERM results are representative of the accuracy of the fitting to a training set. The three methods generally correlate similar descriptors to the experimental values,

Table 4.3: *Descriptors used by the different methods*

Method	**S**	d_1	d_2	d_3	d_4	d_5
RM	0.228	107	278	299	510	1277
MRM	0.228	107	278	299	510	1277
ERM	0.220	37	107	287	527	1277

Since the descriptors are relatively similar it is assumed that they fit the training set similarly well, with the ERM fitting it most effectively.

The simple fact that the RM does not take as long makes it more appropriate for analysis when many datasets are used each requiring linear regression analysis. Using an ERM in this case would require a lot more time and computer effort. An ERM is better when modelling a single or small number of datasets as the increase in accuracy of the method offsets the increase in time taken to obtain the results. In section 6 ERM was used to determine the results for the FIT analysis as only 3 datasets were employed. Whereas in section 7, RM was used since 14 datasets were employed. The extra time taken in obtaining the results for the 14 datasets, if ERM would have been used instead of RM would result in a large and uneconomic time penalty compared to the increase in accuracy of the results in this case.

5 Validation:

5.1 Introduction:

When any QSAR model is proposed, it is fundamental that the model has good predictive ability; this is tested in the validation step. Without a proper validation it is merely a model of similarity between compounds. The validation step is highly important the Organisation for Economic Development and Cooperation (OEDC) stipulated that all QSAR works in publication must have had an adequate validation analysis of its calibration before it can be used in any medical or environmental study[59]. Since the results of the validation step measure the predictive power of the model, it is a valid assumption that this step is in fact the most important in a QSAR study[60].

5.2 External Validation:

An external validation uses a test set which was not used in the training of the model to verify how well the trained model can predict the properties of different molecules[61]. Ideally the test set will present similar results to the calibration set for the optimum number of descriptors. There are many methods available for choosing the test set (n_{test}) from N molecules in the total dataset, some of these are reviewed in Section 7.1. In an ideal situation, with a suitably large number of molecules, a separate validation set is also included in the model. This set differs from the test set, since the validation set can be consulted during the calibration step. The idea behind this is that a calibration that fits the validation set properly, should give similar results for the test set. This validation set scheme is especially important in the method that will be used in K-Means analysis in section 7. Model validation using only data from the training set is called internal validation.

5.3 Internal Validation:

Internal validation of a QSAR model is defined as the analysis of the predictive power of a QSAR model using the data in the training set. The internal validation used in this project is a cross validation study. It is not a substitute for the test set analysis as it is uses molecules in the training set to validate the model. However, it can offer insight into how well the training set would predict values of an external test set. A large increase in S_{loo} (internal validation S) of over 20% compared to S would indicate that the model could not be used in any further analysis, even before looking at the test set results[62].

5.3.1 Cross validation analysis:

Cross validation analysis is the most common method of internal validation analysis used in the literature. It takes the form of "leave one out" (loo) or "leave more out" (lmo) analysis[60]. Leave one out is a validation step that tests the predictive power of the calibration set by removing one of the molecules at random and then recalculating a new model. Obtaining in this way a predicted value for all the molecules; a S_{loo} and a R_{loo}. Huge deviations in the S_{loo} compared to the standard S would be indicative of a model with little predictive power[62]. Leave more out is similar to the loo method but for this analysis a manually selected percentage of the molecules in the training set is taken out instead of only one molecule[9,62]. For example from a training set of 50 compounds, 10% could be removed at random and a new model is calculated for the remaining 45 compounds, predicting with it the removed molecules. This method is more computationally demanding due to the number of different possible sets of 5 o take from 50 compounds.

***Table 5.2**: S, R, S_{loo},R_{loo} and % different between S and S_{loo} for increasing numbers of descriptors d in the model*

d	S	R	S_{loo}	R_{loo}	$\%_{diff}$
1	1.772	0.736	1.820	0.719	2.637
2	1.381	0.852	1.450	0.836	4.759
3	1.261	0.881	1.412	0.850	10.694
4	0.986	0.930	1.114	0.911	11.490

loo analysis starts from the multi variable linear regression QSAR equation,

$$P_{pred} = c_1 + c_2 d_1 + c_3 d_2 \ldots + c_x d_{x-1} \tag{5.1}$$

where P_{pred} is the predicted value for the property being studied and the d are the descriptors chosen to represent the property. A molecule is then removed from the training set at random. The QSAR equation is then recalculated using the new training set but with the same descriptors. This leads to an equation which only differs in the regression constants,

$$P_{loo} = c'_1 + c'_2 d_1 + c'_3 d_2 \ldots + c'_x d_{x-1} \tag{5.2}$$

The P_{loo} is now calculated for the molecule that was removed from the training set. This method continues removing one molecule at a time, calculating its P_{loo} then placing it back in the training set and removing another one to calculate its P_{loo}. These property values are placed against the experimental values on a graph and the points on this graph are analysed using linear regression. The S_{loo} is the standard deviation of the loo points with respect to this line of regression and the R_{loo} is the correlation coefficient of these points with respect to the same line[62]. lmo analysis uses the same method as loo analysis but removes more molecules at each step. This is a much sterner test of the accuracy of the model than loo, as removing more than one molecule changes the whole appearance of the training set. loo undemanding computational complexity compared to lmo analysis means in can be run simultaneously with the linear regression analysis. lmo analysis is generally carried out to achieve confirmation of model accuracy before the publication of a study[62].

In the following section a novel use of loo analysis is presented, the new approach uses R_{loo} to determine the optimum number of descriptors to use in a QSAR model.

6 Determining the optimum number of molecular descriptors to include in a model:

6.1 Introduction:

A fundamental step in the making of a QSAR model is the determination of the optimum number of descriptors (d_{opt}) to include in the QSAR equation. This is a challenging step because as the number of molecular descriptors is increased, the *S* of the training set tends to decrease, since the model becomes closely fitted to the molecules in the set[63]. However, past a particular number of descriptors (*d*), the improved statistical accuracy of the training set comes at the detriment to the statistical accuracy of the test set due to overfitting. (Figure. 6.1.)

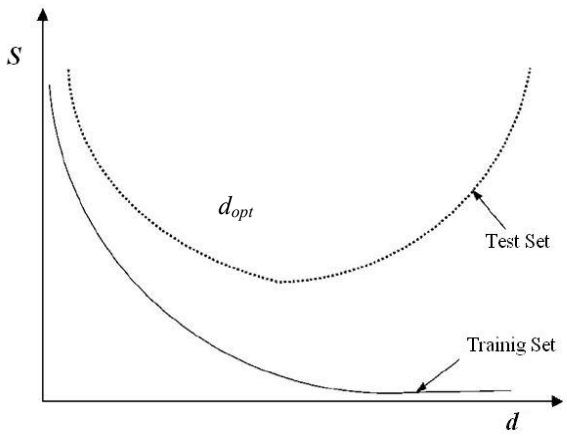


Figure 6.1*: General behavior of S of the training and test set with increasing number of molecular descriptors (d).*

As the number of molecular descriptors is increased from *d = 1,* the *S* of both the training and test sets decreases, due to the fact that almost all molecules can be related by a small number of molecular descriptors. This decrease continues up to a particular value of *d*, (d_{opt}). After this the *S* of the training set continues to decrease but the *S* of test set increases. As the number of descriptors in the model increases the training set becomes very closely fitted to these chosen descriptors. The external test set, which hasn't been involved in the determination of the chosen descriptors, is not then well related to the descriptors chosen with respect to the training set, so the *S* of this test set rises[63]. An example using data taken from a PKA QSAR study is shown in Fig. 6.2. The graph has the same general form as the theoretical graph shown in Fig. 6.1. However in this case the test set has a lower *S* than the training set until the *d* = 6, this is because the size of the test set was very small.

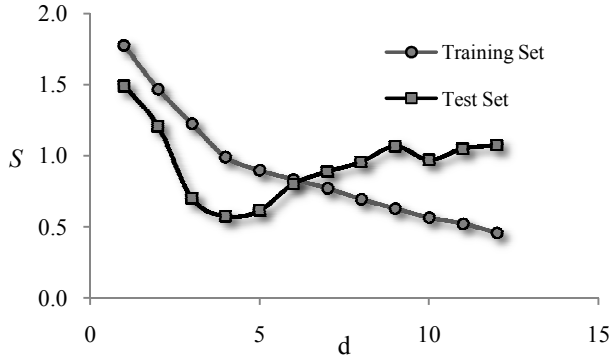


Figure 6.2: *S of the training and test set with increasing d.*

However since the test set for the model is forbidden from being used in the development of QSAR models, these graphs can only be used as a confirmation step after the completion of the model. Consequently it was decided to further develop existing techniques for determining the model dependant d_{opt} in order to make them more

effective. As mentioned beforehand one cannot simply use as many descriptors as possible to achieve a low training set S (S_{train}) as this would give a model with predictive capabilities for the molecules in the training set and not for external molecules. Therefore in this project a development of the Kubinyi equation[64] was used to determine d_{opt}. For all datasets used in this chapter, ERM was used for the calculations.

6.2 Kubinyi FIT equation:

The Kubinyi equation is itself based on the Fisher Ratio[65] which is a statistical test of the accuracy of the linear regression applied to any dataset. The Fisher equation has the form,

$$F_{1,N} = (N-2)\frac{R^2}{1-R^2} \tag{6.1}$$

where N is the number of molecules in the training set and R is the regression coefficient. The equation was found to be too sensitive to changes in low values of N and not sensitive enough to changes in large values of N[63]. Therefore the Kubinyi Fit equation was used in many QSAR research projects as it does not contain these disadvantages. The Kubinyi fit equation is,

$$FIT_R = \frac{R^2(N-d-1)}{(N+d^2)(1-R^2)} \tag{6.2}$$

where d is the number of descriptors used in the model, N is the number of molecules in the training set and R is the regression coefficient.

Plotting the values of FIT versus the values of d should give a graph where the FIT values rise up to a certain value of d, then fall after this point as the increases in R should become too small compared to the increase in d. This first available maximum on the graph is the d_{opt} value.

6.3 VFIT:

Sometimes the FIT values do not form a maximum within a reasonable scope of molecular descriptors, this is mainly because the Kubinyi FIT equation was developed using the less accurate step-wise regression instead of the ERM, that was used in this project[64]. An example is shown here in Fig. 6.3 using RAD.

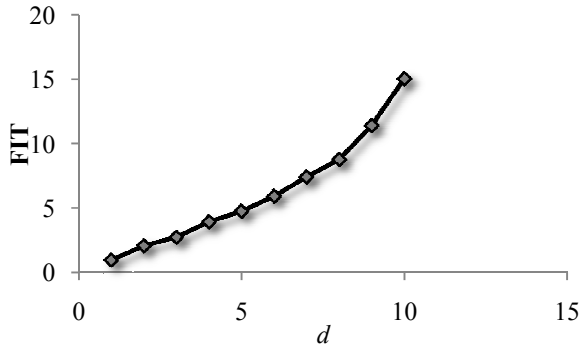


Figure 6.3: *Change in FIT values with increasing number of molecular descriptors.*

Owing to this trend, a development of the Kubinyi equation called VFIT was used during this project. It is the same as the FIT equation except it includes a semi empirical constant k which adds more weight to the value of the number of descriptors d in the FIT equation[66]. k can be any value but in practice is altered in increments of 0.5. This was devised in order to create a maximum for d_{opt}

$$VFIT = \frac{R^2(N-kd-1)}{(N+d^2)(1-R^2)} \tag{6.3}$$

An example of how changing k can lead to a value for d_{opt} using this VFIT equation is demonstrated in Fig. 6.4.

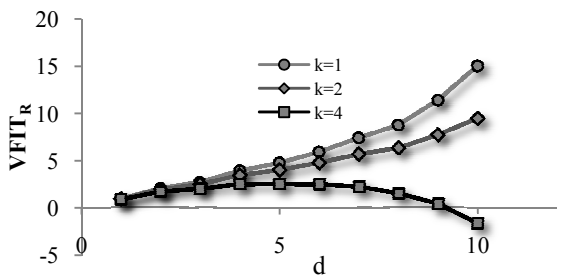


Figure 6.4: *Change of VFIT with different values of k from the RAD dataset*

Here, by altering the values of *k,* a maximum has been achieved at $k = 4$. This corresponds to the results obtained by looking at the test set results for *d,* which show that the test set results for $d = 4$ are the most accurate, thus showing $d = 4$ is the model with the greatest predictive power.

Table 6.1: *Results from ERM of RAD set.*

d	S_{train}	S_{test}
2	0.295	0.499
3	0.235	0.495
4	0.200	0.396
5	0.173	0.555

6.3.1 Development of VFIT using R_{loo} analysis of PKA dataset:

Since the FIT and VFIT equations are used for determining the optimum number of descriptors without looking at the test set, it was decided to use R_{loo} instead of R in the VFIT equation. The theory behind it is that because R_{loo} is a validation parameter - albeit for the training set – it would be more effective at relating the number of descriptors to the test set than R for the training set. Substituting R_{loo} for R in the VFIT equation gives the new equation,

$$VFIT_{loo} = \frac{{R_{loo}}^2(N-kd-1)}{(N+d^2)(1-{R_{loo}}^2)} \tag{6.4}$$

It occurs that $R_{loo} < R$, (Section 5) so therefore the $VFIT_{loo}$ values will be smaller than the standard VFIT values. Table 6.1 shows this relationship for $k = 1$,

Table 6.2: *VFIT vs $VFIT_{loo}$ for different values of d. (k=1, PKA dataset)*

d	VFIT	$VFIT_{loo}$
1	1.125	1.021
2	2.041	1.787
3	3.094	3.099
4	4.689	3.987
5	5.280	3.865
6	5.545	3.879
7	5.806	4.199
8	6.356	4.094
9	7.019	4.434
10	7.792	4.955
11	8.131	4.518
12	9.501	5.389

This data can be plotted to show the difference between the two FIT series,

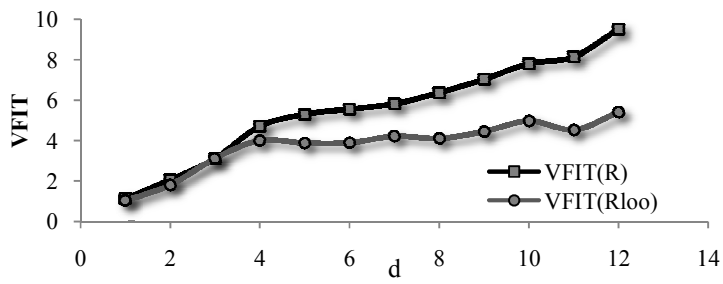


Figure 6.5: *Graphical portrayal of VFIT vs $VFIT_{loo}$ for PKA set*

The graph shows that by placing the R_{loo} into the FIT equation gives a local maximum at *d*=4 even before introducing *k* values, whereas the standard R based FIT increases for the 12 descriptors included in the model with no local maximum. According to Fig. 6.6 the d_{opt} is at *d*=4, however, values of *k* for VFIT have to be increased to obtain this d_{opt}.

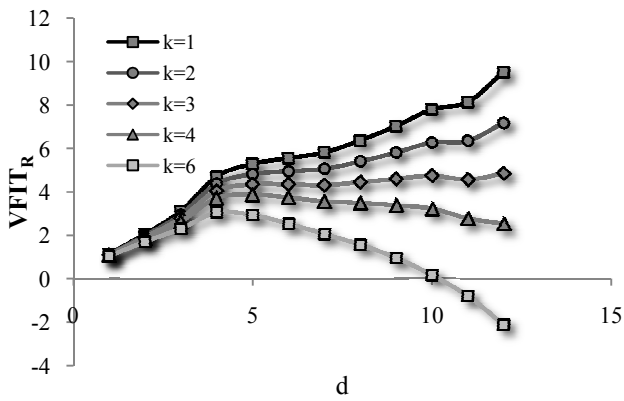


Figure 6.6: *VFIT using R with increasing k values (PKA)*

Only by increasing the value of k to 6 the d_{opt} from VFIT can be achieved. However when increasing *k* up to *k* = 6, maximums are also found for d = 5 when k= 3 and k = 4. This problem does not occur with the $VFIT_{loo}$ for this dataset, which gives d_{opt} = 4, (Fig 6.5). This is supported by the publication, which uses d_{opt} = 4 as the optimum number of descriptors[22].

The different values of d_{opt} that come with different values of *k* in Fig. 6.6 present a problem in the QSAR analysis. Generally the maximum that occurs with the lowest value of k is taken as d_{opt},[64] in this case however this maximum is at *d=5* whereas the the preferable number of descriptors to use in this QSAR model is 4. This would lead to choosing *d*=5 according to VFIT, but this is not as accurate for the test set as a 4 descriptor model as can be verified in Table 6.3.

Table 6.3: *QSAR models for 1 to 6 descriptor (PKA).*

d	S_{train}	R_{train}	S_{test}	R_{test}
1	1.773	0.736	1.486	0.713
2	1.464	0.831	1.210	0.830
3	1.221	0.888	0.700	0.942
4	0.986	0.930	0.571	0.966
5	0.894	0.944	0.611	0.959
6	0.830	0.953	0.800	0.941

This shows the S_{test} for d = 4 is smaller than for d = 5, indicating superior predictive power. The fact that $VFIT_{loo}$ gives this d_{opt} without using any values of k means it can be consider a simpler and more accurate alternative in this dataset to VFIT, which gives different values of d_{opt} depending on the value of k used.

6.3.2 $VFIT_{loo}$ using RAD dataset:

Figure 6.4 shows that the value for k needed to achieve a maximum using VFIT with R was k = 4 for the RAD set, as with PKA. The use of R_{loo} instead or R was investigated to see if improved results occurred.

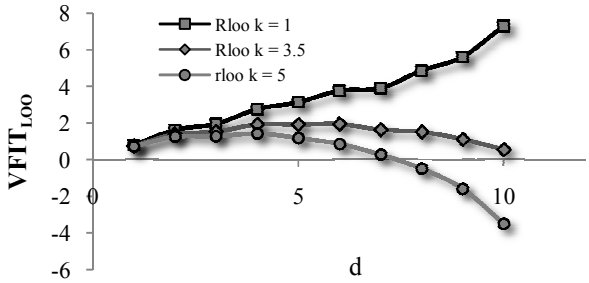


Figure 6.7: *$VFIT_{loo}$ with different values of k for RAD*

There is no maximum at k_{loo} = 1 in Fig. 6.7, neither in the standard R graph. (Fig. 6.4) There is a maximum finally achieved at d = 4 when k_{loo} = 3.5 whereas a maximum was achieved with the same d value when k_R = 4. In this case $VFIT_{loo}$ shows similar results as VFIT, a slight improvement in using R_{loo} could be considered since the local maximum in $VFTI_{loo}$ appears at a lower k value. This adds weight to the idea that $VFIT_{loo}$ is a viable and improved alternative to past R based techniques.

6.3.3 $VFIT_{loo}$ using SINGLET OXYGEN dataset:

One final dataset (SINGLET OXYGEN) was used to further support the improvements of $VFIT_{loo}$ over VFIT for determining d_{opt}.

In this set leaving k=1 gives a graph of VFIT vs. d with no local maximum (Fig. 6.8). Therefore increased values of k would be needed to obtain a local maximum.

Only when k is increased to 2.5 a local maximum a appears in the VFIT graph (Fig. 6.9), this is compared to k = 1 with the $VFIT_{loo}$. Each maximum occurs at d = 6. This value of d_{opt} is vindicated when compared to the QSAR paper in the literature using the SINGLET OXYGEN dataset[21], where d_{opt}=6 also. There is an advantage again of using R_{loo} in the VFIT equation compared to R; a lower value of k is required to obtain the d_{opt}, thus simplifying the selection of d_{opt}.

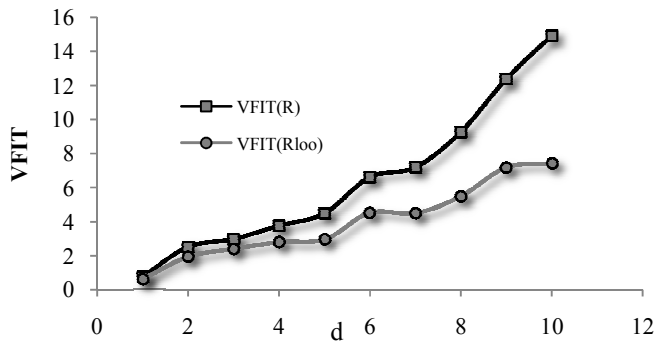


Figure 6.8: *Graphs of VFIT and $VFIT_{loo}$ and how they change with d, k=1, SINGLET OXYGEN dataset*

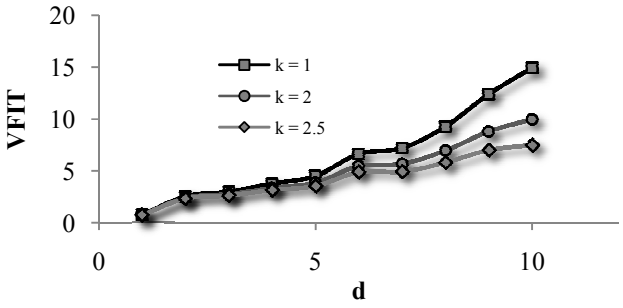


Figure 6.9: *VFIT vs. d with different k values*

6.4 Conclusion:

As shown in each of the three datasets used to test the use of $VFIT_{loo}$, they each give more accurate results than using VFIT. They also give results in concordance with previously published works. Using $VFIT_{loo}$ with the PKA set gives results that greatly exceed the use of VFIT, without needing to change the *k* values. The SINGLET OXYGEN, like the PKA set gives results that match exactly the results from the literature, without needing *k* values in the $VFIT_{loo}$ equation. The RAD set however shows that the $VFIT_{loo}$ technique does, at times, need *k* values to achieve a maximum; but still requires a smaller value of *k* than $VFIT_R$. This proves that using $VFIT_{loo}$ can be used as a stand alone method of determining d_{opt} without needing any reference to VFIT. VFIT could then be used as a test step to determine if it also gives the same d_{opt} as $VFIT_{loo}$. The fact that R_{loo} is a validation correlation coefficient as opposed to a calibration correlation coefficient appears to be essential in leading to a facilitation of calculating the d_{opt}.

Vast amounts of QSAR literature have used Kubinyi FIT and its derivatives to determine d_{opt}. Nevertheless, it is a step that can be omitted if the number of descriptors included in the model is kept constant. Any attempt at omitting this step would only be justified if the accuracy of the results was not affected. Therefore a method of bypassing this descriptor selection stage using K–Means analysis was developed and is presented in the next section.

7 K-Means Cluster analysis:

7.1 Introduction:

There are many methods available to select the training and test sets to carry out a QSAR study. The random method; is a way of choosing a test set when there are no other alternatives; it uses an algorithm that selects a specified number of molecules from the original dataset at random and places them in the test set. The main problem with this method, is that the chosen test set may be bunched together over a narrow activity range and not cover the varied activities of the overall group of molecules. A slightly more effective method would be to sort the activities in ascending order and then to select every x^{th} number for the training set[67]. This would lead to a more representative test set than a random model. The problem with this method is that it still is a manual method based on observation of its activity rather than any reference to its structure. One of QSAR principles is that compounds with similar activities have similar structures, but at times two compounds have similar activities by chance rather than any structural relationship. Therefore a new method of extracting the molecular sets based on the structure and using K-Means analysis is presented. The method will be considered a success if it can match or enhance the results obtained from the linear RM already successfully used in many articles[68].

7.2 K-Means Clustering:

K-Means cluster analysis is a method of partitioning a dataset into *K* mutually exclusive clusters, each representing a certain portion of the aforementioned dataset. It is a geometric clustering algorithm developed by Lloyd in 1982[69] which aims to produce clusters with the smallest possible total distance of the data points to the centre of the cluster (centroid); where the data points in a QSAR study are the molecules in the set. It is an iterative

algorithm that moves points between clusters until the sum of the distance of the cluster points to the centroid cannot be decreased any more, and proceeds as follows:

- The algorithm takes *n* data points and partitions them into *k* clusters, where *k* is inputted manually into the algorithm.
- *k* centroid points are randomly computed and each of the *n* data points is assigned to the nearest centroid point.
- The centroid is then recomputed as to be in the centre of mass of the points assigned to it.
- The data points in each cluster are then reassigned again as the centroid locations have changed, and are allocated to the new closest centroid location.
- These steps are repeated until there are no reassignments possible that will give a lower overall (points to centroids) distance for any of the clusters.

K-Means analysis will be used in this project to first partition the data into clusters based on molecular descriptor values, where each cluster contains similar molecules with similar values. Then these clusters will be used to select the molecule sets used to perform the calculations. A new method of extracting the training and test set based on the structure using K-Means analysis is presented here.

7.3 Partitioning the dataset:

To obtain a training set and test set that is representative of the properties of all the molecules, K-Means clusters have to be developed. Each molecule has a set of molecular descriptors that characterize it, these correspond to the rows of the descriptor matrix, the columns correspond to the values for each molecule of a certain descriptor, and these columns will be used to partition the dataset.

The following data has been partitioned using descriptor 967, chosen at random from the RAD descriptor pool; which belongs to the WHIM descriptor family (Section 3.5.6). It was placed in the K means algorithm and two cluster were partitioned.

The accuracy of the partitioning of the data can be determined by its partition values. These measure the distance of each point in one cluster to points in the neighbouring clusters. These values range from +1 to -1, where a +1 indicates that a point is at the centre of the assigned cluster and -1 indicates that it is assigned erroneously and should be placed in another cluster. It is defined as

$$S_i = (\min(b(i), 2 - a(i))/\max(a(i), \min(b(i), 2)$$

Table 7.1: *Descriptor values and the cluster they are placed in*

Mol Nº	Cluster	Desc. Value	Mol Nº	Cluster	Desc. Value	Mol Nº	Cluster	Desc. Value	Mol Nº	Cluster	Desc. Value
1	1	11.69	14	1	13.342	27	1	11.922	40	1	11.478
2	2	8.935	15	1	14.453	28	2	6.642	41	1	11.691
3	2	8.118	16	1	12.101	29	2	8.14	42	2	7.433
4	2	6.786	17	1	13.57	30	2	5.843	43	1	11.106
5	2	8.33	18	2	9.954	31	2	6.069	44	1	10.837
6	2	10.151	19	2	8.985	32	2	7.492	45	1	16.523
7	1	13.483	20	2	10.291	33	1	11.59	46	1	12.023
8	1	15.616	21	1	12.538	34	2	6.912	47	1	12.658
9	1	13.954	22	2	10.116	35	2	8.433	48	1	13.995
10	1	13.527	23	1	13.121	36	2	5.915	49	1	16.588
11	1	14.283	24	1	12.733	37	2	7.064	50	2	8.74
12	1	13.565	25	2	9.247	38	2	8.435	51	2	9.374
13	1	14.188	26	2	8.224	39	2	6.668	52	2	5.752

Table 7.2: *Partitioning values of the each molecule from the dataset*

Mol. N°	Par Val	Mol. N°	Par Val	Mol. N°	Par Val	Mol. N°	Par Val	Mol. N°	Par Val
1	0.700	12	0.923	23	0.916	34	0.922	45	0.812
2	0.856	13	0.914	24	0.895	35	0.911	46	0.793
3	0.928	14	0.922	25	0.796	36	0.881	47	0.888
4	0.919	15	0.907	26	0.923	37	0.926	48	0.919
5	0.918	16	0.810	27	0.769	38	0.911	49	0.809
6	0.404	17	0.923	28	0.913	39	0.914	50	0.883
7	0.923	18	0.528	29	0.927	40	0.617	51	0.763
8	0.857	19	0.848	30	0.878	41	0.700	52	0.873
9	0.920	20	0.296	31	0.888	42	0.934		
10	0.923	21	0.876	32	0.934	43	0.410		
11	0.912	22	0.428	33	0.666	44	0.192		

where *a(i)* is the average distance from the *ith* point
to the other points in its cluster and *b(i)* is the average distance from the *ith* point to points in the other cluster. This is an equation which takes the minimum value from the denominator out of $b(i)$ or 2 and subtracts $a(i)$, then divides this result by the maximum value from the denominator out of $a(i)$, the minimum value of $b(i)$ or 2. Any points with negative partition values would invalidate the descriptors used to partition the set, and would prevent it from being considered as the partitioning descriptor in further analysis.
Since all the partitioning values in Table 7.1 and Table 7.2 are positive, every point is assigned correctly to each cluster and so this result can be used to extract the molecular sets from the clusters randomly and using a desired percentage for the number of molecules in each set.

7.4 Distance Measurements:

There are a variety of different methods for measuring the distance of each point to the centroid of the cluster it belongs to.

7.4.1 City-Block Distance:

One method of measuring the distance from the points to the centroid is the 'City – Block (Manhattan)' distance measure[8], named as such because it gives the shortest possible distance between two points if the only paths available are segments parallel to the *x* and *y* axis; Comparable to plotting a route between two points in New York.

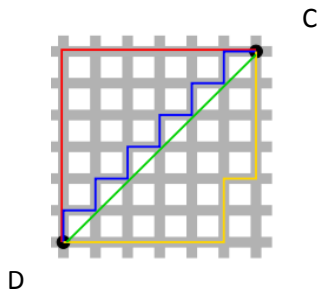


Figure 7.1: *Demonstration of different distance measures*

Here the shortest route between the points $D_{(X,Y)}$ and $C_{(X,Y)}$ is the green line. This is not possible in the 'city-block' measurement so another route must be taken. Here the red blue and yellow vectors are all the same magnitude in 'city-block' geometry.

$$d_1(\mathbf{d},\mathbf{c}) = \sum_{i=1}^{n} |d_i - c_i| \tag{7.2}$$

where **d** and **c** are the horizontal and vertical vectors respectively linking points D (0,0) and C (6,6). The 'city-block' distance in this case is,

$$|0-6|+|0-6|=12$$

7.4.2 Euclidean Distance:

Another distance measuring method is the squared Euclidean method[70], it is the standard geometric distance between two points, such as the green line on the 'city - block' graph. (Fig. 7.1) The equation for the Euclidean distance is as per Pythagarus Theorum,

$$d_1(\mathbf{d},\mathbf{c})^2=\sum_{i=1}^{n}|d_i-c_i|^2 \tag{7.3}$$

Which in the case of Fig. 7.1 is,

$$d_1(\mathbf{d},\mathbf{c})^2=[d_1-d_2]^2+|c_1-c_2|^2=|6-0|^2+|6-0|^2=72$$

taking the square root of each side gives the Euclidean distance of 8.48.

7.4.3 Comparison of distances:

Assuming that the distance measure used is standardised over the *k* clusters, either distance measure can be used to measure the overall distance sum of the points in each cluster to the centroid location. However it is important to note that using different distance measures in the data partitioning step results in different training and test sets. This is because the centroid point is calculated differently for each method. A different centroid location could make some points closer to a different centroid point, hence being assigned to a different cluster. The Euclidean method uses the mean average of all the points in the cluster for each centroid, whereas the 'city - block' method uses a component-wise median of distances as the centroid location. This later centroid location is not a definite point unlike the centroid location in Euclidean distance measurements[8]. A different training and test set inevitably gives different final results, so a measurement must be initially chosen and then maintained for all calculations to be comparable.

Carrying out a basic RM shows that different molecular sets give different results.

Table 7.3: *Differences between the 'Euclidean' and 'City' distance in the RAD; using RM and descriptor 967*

	City - Block		Squared Euclidean	
d	S_{train}	R_{train}	S_{train}	R_{train}
1	0.432	0.744	0.465	0.766
2	0.317	0.875	0.368	0.865
3	0.258	0.922	0.303	0.913

A similar comparison was done on the SOL dataset. Changing the distance measurement for the clusters and carrying out an RM.

Table 7.4: *Difference between the 'Euclidean' and 'City' distance in the SOL; using RM and descriptor 967*

	City - Block		Squared Euclidean	
d	S_{train}	R_{train}	S_{train}	R_{train}
1	1.225	0.735	1.217	0.742
2	1.013	0.830	1.014	0.831
3	0.897	0.871	0.890	0.874

These tables show that changing the distance measure for the clusters gives different results after the selection of the sets from the clusters. The 'city – block' method is more accurate for RAD while the SOL set gives almost the same

results for the two distances. Changing the descriptor number in the dataset also gives different results depending on the descriptor used, for example using the RAD set and changing the descriptor from 967 to 1034 led to more accurate results using the Euclidean Method.

Table 7.5: *Difference between 'Euclidean' and 'City' distance in the SOL; using RM and descriptor1034*

	City Block		Squared Euclidean	
d	S_{train}	R_{train}	S_{train}	R_{train}
1	1.275	0.714	1.254	0.736
2	1.067	0.820	1.020	0.829
3	0.976	0.856	0.912	0.879

7.4.4 Conclusion:

The more conventional Euclidean distance with its defined centroid point appeared more rational and logical hence it was selected for all further K-Means calculations.

However it seems that the accuracy of the obtained results using the sets from each distance measure depends on the property under study and on the descriptor used to partition the data. Since many different properties were to be studied using many different descriptors, either method could have been chosen for the calculations.

7.5 The proposal of a new method for QSAR modeling:

There were previous attempts to use molecular descriptors to partition the dataset in QSAR studies[71,72]. The new proposed methodology employs a new way of obtaining predictions for the molecules properties. The partitioning step uses molecular descriptor values to partition the N molecules into *k* clusters, as was explained in section 7.3. Each descriptor has a different value for every compound, which is used independently to partition the data. A training, validation and test set are chosen from the clusters.

Table 7.6: Average S_{test} for d = 1-3 of different datasets using RM with different percentages for the molecular sets.

Dataset	**TOX**			**PERM**		
$N_{train+val}$ (%)	60	70	80	60	70	80
S_{test}	0.89	0.70	0.91	0.74	0.57	0.59
Dataset	**SOL**			**RAD**		
$N_{train+val}$ (%)	60	70	80	60	70	80
S_{test}	1.56	1.25	1.42	0.56	0.53	0.45

7.5.1 Partitioning Method:

An algorithm was developed that makes use of the K-Means cluster analysis already present in MATLAB. (Section 7.2) The new algorithm developed for this project proceeds uses the following steps;

- A number of clusters *k* is manually inputted into the algorithm.
- Each descriptor from the total descriptor matrix, from $d = 1$ to $d = D$ is placed into the original K-Means algorithm.
- Numerical values are assigned to different clusters depending on their magnitude.
- The algorithm finishes after creating a new descriptor matrix (M_{clust}) that contains a cluster value corresponding to each descriptor.

In addition, every descriptor column has a partitioning value (Table 7.2). Any descriptor with a negative partition value was ignored. Since negative partition values indicate that the data point is not placed in the correct

cluster. Hence, when extracting the training, validation and test sets from each cluster, molecules could be selected for the wrong set, thus damaging the reliability and accuracy of the results.

7.5.2 Constitution of the training, validation and test sets (molecular sets):

It is required to choose a percentage of molecules for the training, validation and test sets. A ratio of N_{train}=N_{val} was fixed, because both training and validation sets are used to select the model; it is therefore valid to assume that they are equally important. Using a higher percentage of compounds in the training and validation sets would increase accuracy of these sets, at the expense of accuracy of the test set; thus invalidating the predictive power of the model.

After the data are partitioned into clusters, the molecular sets are randomly extracted from each cluster proportionally to the number of molecules in them. For example if N_{train}=40%, 40% of the molecules would be taken out of each cluster to form the training set.

Table 7.6 shows that $N_{train+val}$=70% is the most accurate number of molecules in each set for the larger sets of molecules (TOX, PERM, SOL). However when the number of molecules in the dataset is decreased, putting more molecules in the training and validation set (80%) increases the accuracy of the model (RAD). This occurs because with fewer overall dataset molecules, there is less capacity to find patterns between the structures, hence more molecules would have to be used in the training set to counteract this. An appropriate training set fitting should produce a model with adequate results in the test set. The results are shown in table 7.6, from which it follows that the best options are; $N_{train+val}$=70% for datasets with more than 60 molecules; and $N_{train+val}$=80% for datasets such as RAD with fewer than 60 molecules,

7.6 Novel Method description:

The novel method proceeds via the scheme shown in Fig. 7.2. For each dataset the molecules are partitioned into *k* clusters; each cluster has a certain number of molecules that depends on the values of the descriptor used to partition the data. Then a training set, validation set and test set are randomly taken out of each cluster, complying with the selected percentage of training and validation compounds. Each cluster therefore has a training, validation and test set associated with.

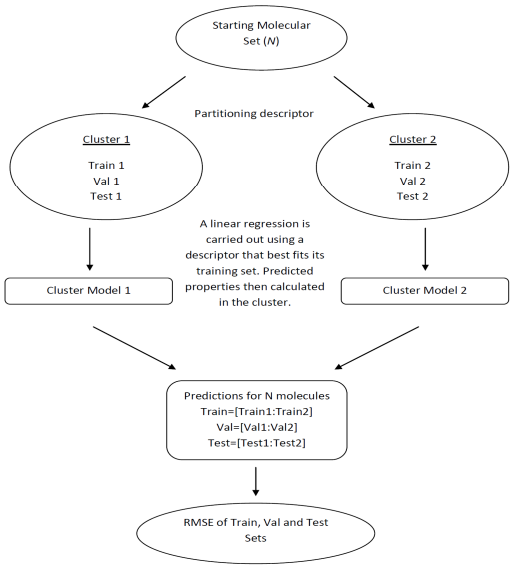


***Figure** 7.2: Flowchart showing the evolution of the new method*

Linear regression is carried out independently in each cluster, to fit the training set molecules to the descriptors from the total descriptor matrix. For each cluster only one descriptor is used to train its model. Then the Root Mean Square Error (RMSE) is calculated for every cluster validation and test sets. RMSE is used, since the number of descriptors selected does not influence the outcome of the fitting, contrary to the use of *S* (section 4.3.3). The RMSE equation is very similar to the *S* equation,

$$RMSE = \sqrt{\frac{\sum(x_i-\mu)^2}{N}} \tag{7.4}$$

The results are then organized into a table that shows in the first row, the lowest percentage difference between the training and validation set, and the corresponding descriptors. And in the final row, the highest percentage difference between the two. For each method the RMSE of training, validation and test sets are used to contrast the results obtained with the RM.

The best result from the validation set is taken forward as the optimum because in theory as the validation set is a pseudo test set, good correlation between the training and validation set, should results in a similar good correlation between the training and test set.

7.6.1 Choosing the value of *k:*

Selecting the number of clusters to partition the data is a vital step in the novel method development. The idea is to partition the data into *k* clusters; that give the best possible predictive power of the model. To determine the best number of clusters to partition the data, different values of *k* were tested. The results were presented as the average RMSE for the training, validation and test sets for all descriptor that gave only positive partition values. Three datasets were used to deduce the optimum number of clusters.

Table 7.7: *Average RMSE of three different dataset sets using 3 different values of k.*

K	Dataset	$RMSE_{train}$	$RMSE_{val}$	$RMSE_{test}$
2	PERM	0.074	0.306	0.502
3		0.050	0.380	0.542
4		0.033	0.412	0.575
2	THIA	0.552	2.086	1.128
3		0.441	2.696	2.139
4		0.329	2.841	2.618
2	PKA	0.523	1.127	1.592
3		0.324	1.625	1.601
4		0.310	1.451	2.468

7.6.1.1 Conclusion:

Table 7.7 shows that the predictive power of the test set is much higher when the data is partitioned into 2 clusters rather than 3 or 4. Contrastingly the $RMSE_{train}$ is much lower when *k* is increased, this occurs because when the data is partitioned into 3 or 4 clusters, each cluster contains fewer molecules than when $k = 2$. Since each cluster contains similar molecules, there is likely to be a strong linear relationship between the training sets for each cluster. This would lead to over fitting of the model for each cluster, giving very accurate results for the training set in detriment of the results for the test set. The same conclusion can be drawn from the results for the validation set; they also become less accurate as *k* is increased. Furthermore at times it is impossible to partition smaller datasets into more than 2 clusters as one cluster could be left empty, as is the case when attempting to partitioning RAD into four clusters. This would have hindered the scale of the developed method if $k > 2$ was chosen as the optimum number of partitioning clusters.

It was therefore decided to partition all datasets into two clusters for optimum predictive power before carrying on with the novel method.

7.6.2 Comparison with existing methods:

The accuracy of the new QSAR method using just one descriptor in each cluster to train the model, was compared with the standard RM model. In order to consider the new methodology a valid alternative to the linear RM methods, it must give similar results to RM even when an optimum number of descriptors in RM have been selected.

Similar results would mean that the developed model could be used to develop a QSAR model, without the extra complications of choosing the correct number of descriptors. Since RM uses *S* to measure the results; they must be converted to RMSE for comparison. The clusters that correspond to the partitioning descriptor which gives the lowest percentage difference of RMSE between the training and validation set; were used to provide the training validation and test set for the RM analysis.

The RM (Section 4.3.3) is used to provide a prompt, accurate model using multiple linear regression. In this case rather than the test set, the validation set was used as the pseudo test set in the algorithm. This provides a matrix of S and R values for the training set and the validation set using different values of *d*. The optimum results for each value of *d* are then used for the final comparison. As there are sometimes many models for each different value of *d* a criteria for its selection was developed. The optimum model would be the one with the lowest percentage difference between *S* of the training set and the *S* of the validation set. However models were rejected if they presented a linear dependence between the descriptors (C_{max}) greater than 98%. Linear dependence between descriptors damages the accuracy of the model; choosing just one of these linear dependant descriptors would give the same result as taking both.

Table 7.8: *Different results obtained using RM with d=6*

d	$\mathbf{S_{train}}$	$\mathbf{S_{val}}$
1	0.47	0.68
2	0.41	0.65
3	0.39	0.60
4	0.38	0.59
5	0.34	0.53
6	0.33	0.53
7	0.32	0.54

As an example, Table 7.8 shows the results for $d = 6$ using RM, result N° 2, has the lowest $\%_{diff}$ between *S* and S_{val}, in addition, it has an acceptable correlation between the descriptors. The sixth result is unacceptable as it has a correlation between the descriptors of over 98%. This approach was carried out for values of *d*=1 to *d*=7 for all datasets.

To compare the results RMSE for the training, validation and test sets is also calculated.

Table 7.9: *S_{train} and S_{val} as a function of d for RM*

d	Descriptors used and their names													
1	1498	LOGKP												
2	175	1498	IC1	LOGKP										
3	116	177	1498	X3v	SIC1	LOGKP								
4	50	492	1214	1498	AAC	GATS1v	RTv+	LOGKP						
5	7	1214	1302	1498	1504	Mv	RTv+	nNH2	LOGKP	Nahalket				
6	7	1214	1297	1302	1498	1504	Mv	RTv+	nCOH	nNH2	LOGKP	Nahalket		
7	7	1214	1297	1299	1302	1498	1504	Mv	RTv+	nCOH	nCO	nNH2	LOGKP	Nahalket

The descriptors that correspond to these values of S_{train} and S_{val} are shown in Table 7.10.
The $RMSE_{test}$ was then calculated from S_{test} by using the relationship between the two equations, and presented in Table 7.11. The best model for *d* is deduced using the *S* values from the RM (table 7.11), uniform criteria was needed to ensure fairness when comparing the linear RM against the novel method.The following criteria was used to select the optimum number of descriptors from RM, they are shown in order or decreasing importance;
N_{train} / d must be greater than 10

Lowest S_{val}

If similar S_{val}, lower S_{train}

If similar S_{val} and S_{train}, then the model with the fewer number of descriptors is preferable

When the optimum number of *d* is selected, the RMSE for the training, validation and test set for this value of *d* is compared to the RMSE of the developed method.

Table 7.10: *Descriptors that correspond to the results obtained in Table 7.9*

d	$RMSE_{train}$	$RMSE_{val}$	$RMSE_{test}$	$RMSE_{(val-train)}$	$\%_{dif}$	cxmax
1	0.47	0.68	0.94	0.21	31.07	0.00
2	0.40	0.65	0.93	0.25	37.96	0.41
3	0.38	0.59	0.81	0.21	34.99	0.57
4	0.37	0.58	0.84	0.21	36.57	0.65
5	0.34	0.52	0.70	0.18	34.47	0.82
6	0.32	0.51	0.69	0.19	36.93	0.82
7	0.31	0.53	0.74	0.22	41.08	0.82

Table 7.11: *RMSE that correspond to models shown in table 7.10*

d	$RMSE_{train}$	$RMSE_{val}$	$RMSE_{test}$	$RMSE_{(val-train)}$	$\%_{dif}$	cxmax
1	0.47	0.68	0.94	0.21	31.07	0.00
2	0.40	0.65	0.93	0.25	37.96	0.41
3	0.38	0.59	0.81	0.21	34.99	0.57
4	0.37	0.58	0.84	0.21	36.57	0.65
5	0.34	0.52	0.70	0.18	34.47	0.82
6	0.32	0.51	0.69	0.19	36.93	0.82
7	0.31	0.53	0.74	0.22	41.08	0.82

7.7 Results:

The K-Means partitioning method was tested for its robustness against a total of 14 datasets, 9 with greater than 60 molecules in the dataset, and 5 with fewer than 60 molecules.

7.7.1 Greater than 60 molecules:

Table 7.12: *Comparison of the RM with the K-Means cluster analysis method for molecular sets with greater than 60 molecules*

SOL							
N_{train}	57		N_{val}	59		N_{test}	50
K – Means				RM			
$d_{partition}$	$RMSE_{train}$	$RMSE_{val}$	$RMSE_{test}$	d_{opt}	$RMSE_{train}$	$RMSE_{val}$	$RMSE_{test}$
127	0.94	1.28	0.99	2	0.81	1.24	0.98
HIV							
N_{train}	44		N_{val}	45		N_{test}	39
K – Means				RM			
$d_{partition}$	$RMSE_{train}$	$RMSE_{val}$	$RMSE_{test}$	d_{opt}	$RMSE_{train}$	$RMSE_{val}$	$RMSE_{test}$
408	0.55	0.92	1.05	4	0.43	0.84	1.00
FLUOR							
N_{train}	40		N_{val}	41		N_{test}	35
K – Means				RM			
$d_{partition}$	$RMSE_{train}$	$RMSE_{val}$	$RMSE_{test}$	d_{opt}	$RMSE_{train}$	$RMSE_{val}$	$RMSE_{test}$
1021	0.93	1.33	1.20	2	0.48	1.27	1.21
TOX							
N_{train}	164		N_{val}	165		N_{test}	141
K - Means				RM			
$d_{partition}$	$RMSE_{train}$	$RMSE_{val}$	$RMSE_{test}$	d_{opt}	$RMSE_{train}$	$RMSE_{val}$	$RMSE_{test}$
36	0.40	0.64	0.83	5	0.34	0.52	0.70
MES							
N_{train}	34		N_{val}	36		N_{test}	30
K - Means				RM			
$d_{partition}$	$RMSE_{train}$	$RMSE_{val}$	$RMSE_{test}$	d_{opt}	$RMSE_{train}$	$RMSE_{val}$	$RMSE_{test}$
374	0.41	0.35	0.50	1	0.46	0.33	0.46
PERM							
N_{train}	28		N_{val}	28		N_{test}	14
K - Means				RM			
$d_{partition}$	$RMSE_{train}$	$RMSE_{val}$	$RMSE_{test}$	d_{opt}	$RMSE_{train}$	$RMSE_{val}$	$RMSE_{test}$
1075	0.09	0.31	0.46	2	0.07	0.33	0.53
THIA							
N_{train}	30		N_{val}	30		N_{test}	15
K - Means				RM			
$d_{partition}$	$RMSE_{train}$	$RMSE_{val}$	$RMSE_{test}$	d_{opt}	$RMSE_{train}$	$RMSE_{val}$	$RMSE_{test}$
506	0.77	1.51	1.42	2	0.71	1.42	1.53
PKA							
N_{train}	31		N_{val}	33		N_{test}	16
K - Means				RM			
$d_{partition}$	$RMSE_{train}$	$RMSE_{val}$	$RMSE_{test}$	d_{opt}	$RMSE_{train}$	$RMSE_{val}$	$RMSE_{test}$
175	0.42	0.57	0.66	1	0.47	0.73	0.64
GABA							
N_{train}	31		N_{val}	31		N_{test}	16
K - Means				RM			
$d_{partition}$	$RMSE_{train}$	$RMSE_{val}$	$RMSE_{test}$	d_{opt}	$RMSE_{train}$	$RMSE_{val}$	$RMSE_{test}$
1112	0.50	1.06	1.25	2	0.46	1.03	0.68

7.7.2 Fewer than 60 molecules:

Table 7.13: *Comparison of the RM with the K-Means cluster analysis method for molecular sets with fewer than 60 molecules*

RAD							
N_{train}	21		N_{val}	21	N_{test}		10
K - Means				RM			
$d_{partition}$	$RMSE_{train}$	$RMSE_{val}$	$RMSE_{test}$	d_{opt}	$RMSE_{train}$	$RMSE_{val}$	$RMSE_{test}$
107	0.48	0.43	0.34	1	0.48	0.43	0.34
NAFT							
N_{train}	17		N_{val}	16	N_{test}		8
K - Means				RM			
$d_{partition}$	$RMSE_{train}$	$RMSE_{val}$	$RMSE_{test}$	d_{opt}	$RMSE_{train}$	$RMSE_{val}$	$RMSE_{test}$
913	0.81	1.08	1.25	1	0.81	1.08	1.25
MALAR							
N_{train}	8		N_{val}	9	N_{test}		5
K - Means				RM			
$d_{partition}$	$RMSE_{train}$	$RMSE_{val}$	$RMSE_{test}$	d_{opt}	$RMSE_{train}$	$RMSE_{val}$	$RMSE_{test}$
854	0.19	0.36	0.68	1	0.19	0.36	0.68
45 compounds selected from GABA							
N_{train}	18		N_{val}	18	N_{test}		9
K - Means				RM			
$d_{partition}$	$RMSE_{train}$	$RMSE_{val}$	$RMSE_{test}$	d_{opt}	$RMSE_{train}$	$RMSE_{val}$	$RMSE_{test}$
434	0.55	0.62	1.35	1	0.55	0.62	1.35
40 molecules selected from PERM							
N_{train}	15		N_{val}	17	N_{test}		8
K - Means				RM			
$d_{partition}$	$RMSE_{train}$	$RMSE_{val}$	$RMSE_{test}$	d_{opt}	$RMSE_{train}$	$RMSE_{val}$	$RMSE_{test}$
43	0.09	0.14	0.14	1	0.09	0.14	0.14

If fewer than 60 molecules are used the method differs slightly. When the data is partitioned into the two clusters, the molecular sets are extracted and grouped together as a total training set, validation set and test set, rather than separate independent sets in each cluster. The predictions are then obtained from just the one training set, rather than two training sets. This is because in each cluster the training set would contain insufficient number of molecules and could be overfitted by one descriptor; this would come at the expense of inaccuracy in the validation and test set.

7.8 Conclusion:

The aim of the developed technique using K-Means clustering was to provide a new method for QSAR analysis that could give similar or better results for the predictive potential ($RMSE_{test}$) of the dataset than the existing linear methods. It would have the advantage of not needing to choose the amount of descriptors to place in the model; (section 6) resulting in a simpler modelling technique. The $RMSE_{train}$ and $RMSE_{val}$ are often poorer in the novel method, compared to the RM as more descriptors are used to train the model in the RM. These however are of secondary importance to $RMSE_{test}$, which shows the true predictive potential of the model. For each dataset with greater than 60 molecules the developed method gives a similar predictive potential ($RMSE_{test}$) compared to the linear RM method. In the case of the PERM, FLUOR and THIA it gives an improved $RMSE_{test}$ compared to the RM method. The only noticeable deviation from the similar accuracy is the GABA set; where the RM works better, possibly due to the fact that more descriptors are used in d_{opt} compared to the developed method; even though this

doesn't seem to affect the other sets. In the datasets with fewer than 60 compounds there is complete agreement between the RM and the developed method. For each set the $RMSE_{train}$, $RMSE_{val}$ and $RMSE_{test}$ are the same. This is because the method is effectively the same, choosing one descriptor to train the same set of molecules in both the RM and the developed method.

The results for each dataset show that the novel method is in fact a very viable alternative to the standard linear QSAR methods that have dominated the field in the last decades[73]. The idea of not requiring selecting the amount of descriptors to use in the model greatly decreases the time taken for each QSAR study. In addition it would eliminate the possible inaccuracies that occur when choosing the wrong number of molecular descriptors in the model(Section 6).

The new method also reduces the computational time when the size of the total descriptor pool (D) is increased without affecting the accuracy of the results.

The K-Means clustering method greatly facilitates the QSAR process as it misses out the descriptor selection step. By making QSAR methodology easier to carry out, more studies could be completed; therefore more information can be mapped and more properties can be investigated. Many different fields benefit from QSAR studies including pharmaceutical [74] and agricultural [75] sectors so more QSAR studies would aid development in these areas and the wider economy that they are a part of, representing a benefit to modern society.

References:

[1] D. Rouvray, and D. Bonchev, *Chemical Graph Theory: Introduction and Fundamentals,* Volume **1**, P 300 Tunbridge Wells: Abacus Press, 1991.

[2] L. P. Hammett, *J. Am. Chem. Soc.,* **59**, p 96, 1937.

[3] R. W. Taft, *J. Am. Chem. Soc.,* **74**, p 2729, 1952.

[4] P. R. Duchowicz, A. G. Mercader, F. M. Fernandez, and E.A. Castro, *Chemometr. Intell. Lab. Syst.,* **90(2)**, p 97, 2008.

[5] P. R. Duchowicz, A. Talevi, L. E. Bruno-Blanch, and E. A. Castro, *Bioorg. Med. Chem.*, **16(17)**, p 7944, 2008.

[6] J. D. Mckinney, A. Richard, C. Waller, M. C. Newman, and F. Gerberick, *Toxicological Sci.,* **56**, p 8, 2000.

[7] R. Todeschini, and V. Consonni, *Handbook of Molecular Descriptors.* p 688, Weinheim: Wiley – VCH. 2002.

[8] J. Jaworska, N. Nikolova-Jeliazkova, and T. Aldenberg, *ATLA,* **33**, p 445, 2005.

[9] P. Gramatica, *Qsar Comb. Sci.*, **26**, p 694, 2007.

[10] A. J. Scott, and M. Knott, *Biometrics*, **30**, p 507, 1974.

[11] R. M. V. Abreu, I. C. F. R. Ferreira, and M. J. R. P.Queiroz, , *Eur. J. Med. Chem.*, **44(5)**, p 1953, 2009.

[12] T. Katsube, H. Tabata, Y. Phta, Y. Yamasaki, E. Anuurad, K.Shiwaku, and Y. Yamane, *J. Agric. Food Chem.,* **52**(8), p 239, 2004.

[13] H. X. Liu, R. S. Zhang, X. J. Yao, M. C. Liu, Z. D. Hu, and B. T. Fan *J. Chem. Inf. Comput. Sci.*, **43**(4), p 1290, 2003.

[14] D. Plavšić, S. Nikolić, N. Trinajstić, and Z. Mihalić, *J.Math. Chem.,* **12(1)**, p 235, 1993.

[15] J. L. Rodgers, and W. A. Nicewander, *Am. Stat.,* **42(1)**, p 59, 1988.

[16] V. Ravichandran, V. K. Mourya, and R. K., Agrawal, *J. Enzyme. Inhib. Med. Chem*, **26(2)**, p 291, 2011.

[17] A. G. Mercader, and A. B. Pomilio, *Eu. J. Med. Chem*, **45(5)**, p 1727, 2010.

[18] P. R. Duchowicz, A. Talevi, L. E. Bruno-Blanch, and E. A. Castro, *Bioorg. Med. Chem.,* **16(17)**, p 7947, 2008.

[19] P. R. Duchowicz, and M. A. Ocsachoque, *QSAR & Comb. Sci.* **28(3)**, p 281, 2009.

[20] P. R. Duchowicz, M. Fernandez, J. Caballero, E. A. Castro, and F. M. Fernandez, *Bioorg. Med. Chem.,* **14(17)**, p 5878, 2006.

[21] A. G. Mercader, P. R. Duchowicz, F. M. Fernandez, E. A. Castro, F. M. Cabrerizo, and A. H. Thomas, *J. Mol. Graph. Model.,* **28(1)**, p 14, 2009.

[22] S. Son, and B. A. Lewis, *J. Agric. Food. Chem.,* **50(3)**, p 470, 2002.

[23] A. G. Mercader, P. R. Duchowicz, M. A. Sanservino, F. M. Fernandez, and E. A. Castro, *J. Fluor. Chem.,* 1**28(5)**, p 485, 2007.

[24] I. O. Edafiogho, C. N. Hinko, H. Chang, J. A. Moore, D. Mulzac, J. M. Nicholson, and K. R. Scott, *J. Med. Chem.,* **35(15)**, p 5266, 1992.

[25] I. O. Edafiogho, V. V. Ananthalakshmi, and S. B., Kombian, *Bioorg. Med. Chem.,* **14(15)**, p 5267, 2006.

[26] K. Sudo, Y. Matsumoto, M. Matsushima, M. Fujiwara, K. Konno, K. Shimotohno, S. Shigeta, and Y. Yokota, *Biochem. Biophys. Res. Comm.,* **238(2)**, p 643, 1997.

[27] A. G. Mercader, M. Goodarzi, P.R. Duchowicz, F. M. Fernandez, and E. A. Castro, *Chem. Biol. Drug Des.,* **76(5)**, p 434, 2010.

[28] P.R. Duchowicz, M. G. Vitale, E. A. Castro, J. C. Autino, G. Romanelli, D. O. Pand Bennardi, *Eur. J. Med. Chem.,* **43(8)**, p 1594, 2007.

[29] M. Karelson, G. Karelson, T. Tamm, I. Tulp, J. Janes, K. Tamm, A. Lomaka, D.Savchenko, and D. Dobchev, *ARKIVOC*, **II**, p 219, 2009.

[30] C. A. Molyneaux, M. Krugliak, H. Ginsburg, K. Chibale, *Biochem. Pharmacol.,* **71**, p 61, 2005.
[31] P.R. Duchowicz, E. L. Bonifazi, L. G. Leon, M. C. Hernandez, J. M. Padron, G. Burton, E. A. Castro, and R. I. Misico, *XVII simposio Nacional de Química Orgánica, Sociedad Argentina de Investigación en Química Orgánica, Mendoza, 15 al 18 de Noviembre de* 2009.
[32] C. Hansch, and F. Toshio, *J. Am. Chem. Soc.,* **86(8)**, p 1616, 1964.
[33] I. V. Tetko, J. Gasteiger, R. Todeschini, A. Mauri, D. Livingstone, P. Ertl, V. A. Palyulin, E. V. Radchenko, N. S. Zefirov, A. S. Makarenko, V. Y. Tanchuk, and V. V. Prokopenko, *J. Comput. Aid. Mol. Des.*, **19**, p 453, 2005.
[34] HYPERCHEM, 6.03. (Hypercube): http://www.hyper.com.
[35] A. Beck, M. N. Bleicher, and A. D. Crowe, *Excursions Into Mathematics: The Millennium Edition.*, p 500, Wellesley: Massachusetts, A K Peters, Ltd. 2000.
[36] L. Euler, *Commentarii Academiae Imperialis Petropolitanea*, **8**, p 128, 1736.
[37] J. Dugundji, and I. Ugi, *Comput. Chem.* **39**, p 19, 1973.
[38] K. M. Khoda, Y. Liu, and C. Storey, *J. Optim. Theory. Appl.*, **75(2)**, p 345, 1992.
[39] S. Mcqueen – Mason, and D. J. Cosgrove, *Proc. Natl. Acad. Sci. USA,* **91(14)**, p 6574, 1994.
[40] G. B. Moreau, and P. Broto, *Nouv. J. Chim.,* **4**, p 757, 1980.
[41] F. R. Burden, *Chem. Inf. Comput. Sci.,* **29(3)**, p 225, 1989.
[42] J. W. Reynolds, and L . C. Cusachs, *J. Chem. Phys.,* **43(10)**, p 160, 1964.
[43] J. Galvez, *J. Chem. Inf. Comput. Sci*., **34**, p 520, 1994.
[44] J. Galvez, *J. Chem. Inf. Comput. Sci*., **35**, p 272, 1995.
[45] I. Rios – Santamarina, R. Garcia – Domenech, J. Galvez, J. Cortijo, Santamaria, E. Pand Marcillo, *Bioorg. Med. Chem. Lett.,* **8**, p 477, 1998.
[46] J. P. Stewart, *J. Comput. Chem.* **11(4)**, p 543, 1990.
[47] A. Hocquet, and M. Langgard, *J. Molecular Modelling,* **4**(3), p 94, 1998.
[48] V. Consonni, R. Todeschini, and M. Pavan, *J. Chem. Inf. Model.*, **42**, p 693, 2002.
[49] M. Randic, *New J. Chem.*, **19, p** 781, 1995.
[50] C. Pieri, M. Marra, F. Moroni, R. Recchioni, and F. Marchselli, *Life Sciences,* **55**(15), p 272, 1994.
[51] J. Gasteiger, J. Sadowski, J. Schuur, P. Selzer, L. Steinhauer, and V. Steinhauer, *J. Chem. Inf. Comput. Sci.,* **36**(5), p 1030, 1996.
[52] M. P. Gonzalez, C. Teran, M. Teijeira, and A. M. Helguera, *Eur. J. Med. Chem.*, **41**(1), p 58, 2006.
[53] R. Todeschini, and P. Gramatica, *Quantity Structure Activity Relationships,* **16**(2), p 113, 1997.
[54] P. Gramatica, M. Corradi, and V. Consonni, *Chemosphere,* **41,** p 763, 2000.
[55] J. E. Dayhoff, and J. M. Deleo, *Cancer,* **91**(8), p1615, 2001.
[56] H. M. Fatemi, and S. Gharaghani, *Bioorg. & Med. Chem.*, **15**(24), p 7746, 2007.
[57] P.R. Duchowicz, E. A. Castro, and F. M. Fernandez, *MATCH Commun. Math. Comput. Chem.,* **55**, p 179, 2006.
[58] A. G. Mercader, P.R. Duchowicz, F. M. Fernandez, and E. A. Castro, *Chemometr. Intell. Lab. Syst.*, **92**(2), p 138, 2008.
[59] G. Kateman, and J. R. M. Smits, *Analytica Chimica Acta,* **277(2)**, p 179, 1993.
[60] S. S. So, S. P. Van Helden, V. J. Van Geerestein, and M. Karplus, *J. Chem. Inf. Model.,* **40(3)**, p 762, 2000.
[61] J. S. Jaworska, M. Comber, C. Auer, and C. Van Leeuwan, *J. Environ. Health Persp.,* **111**, p 1358, 2003.
[62] A. Tropsha, Gramatica, and V. K. Pand Gombar, *QSAR Comb. Sci.*, **22**, p 69, 2003.
[63] A. Golbraikh, and A. Tropsha, *J. Mol. Graphics Model.*, **20(4)**, p 269, 2002.
[64] V. Consonni, D. Ballabio, and R. Todeschini, *J. Chem. Inf. Model.,* **49(7)**, p 1669, 2009.
[65] D. Hawkins, *J. Chem. Inf. Comput. Sci.,* **43**, p 579, 2004.
[66] H. Kubinyi, *J.Med. Chem.,* **20(5)**, p 625, 1977.
[67] R. A. Fisher, *Proc. Royal Soc. Edinb.,* **42**, p 321, 1922.
[68] A. G., Mercader, P.R. Duchowicz, F. M. Fernandez, E. A. Castro, and E. Wolcan, *Chem. Phys. Let.,* **462(4)**, p 354, 2008.
[69] A. Golbraikh, and A. Tropsha, *J. Comput. Aid. Mol. Des.,* **16**, p 357, 2002.
[70] A. H. Morales, P.R. Duchowicz, M. A. Perez, E. A. Castro, M. N. D. Cordeiro, and M. P. Gonzalez, *Chemometr. Intell. Lab. Syst.,* **81 (2)**, p 180, 2006.
[71] S. P. Lloyd, *IEEE Trans. Inf. Theory,* **28(2)**, p 129, 1982.
[72] P.E. Danielsson, *Comp. Graph. Image Process.,* **14(3)**, p 227, 1980.
[73] S. Liu, C. S. Yin, Z. L. Li, S. X. Cai, *J. Chem. Inf. Comput. Sci.,* **41**(2), p 321, 2001.
[74] R. W. Stanforth, E. Kolossov, and B. Mirkin, *QSAR Comb. Sci.,* **26**, p 837, 2007.
[75] L. T. Leonard, and K. Roy, *QSAR Comb. Sci,* **25(3)**, p 235, 2006.