



University
of Glasgow

Sneddon, Duncan J.M. (2010) *Statistical analysis of crystallographic data*. PhD thesis.

<http://theses.gla.ac.uk/1683/>

Copyright and moral rights for this thesis are retained by the author

A copy can be downloaded for personal non-commercial research or study, without prior permission or charge

This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the Author

The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the Author

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given

Statistical Analysis of Crystallographic Data



University of Glasgow

Duncan J. M. Sneddon

Doctor of Philosophy in Chemistry

Department of Chemistry
University of Glasgow

September 2009

©Duncan JM Sneddon, September 2009

Abstract

The Cambridge structural database (CSD) is a vast resource for crystallographic information[1]. As of 1st January 2009 there are more than 469,611 crystal structures available in the CSD. This work is centred on a program *dSNAP* which has been developed at the University of Glasgow [10]. *dSNAP* is a program that uses statistical methods to group fragments of molecules into groups that have a similar conformation. This work is aimed at applying methods to reduce the number of variables required to describe the geometry of the fragments mined from the CSD.

To this end, the geometric definition employed by *dSNAP* was investigated. The default definition is total geometries which are made up of all angles and all distances, including all non-bonded distances and angles. This geometric definition was investigated in a comparative manner with four other definitions. There were all angles, all distances, bonded angles and distances and bonded angles, distances and torsion angles. These comparisons show that non-bonded information is critical to the formation of groups of fragments with similar conformations.

The remainder of this work was focused in reducing the number of variables required to group fragments having similar conformations into distinct groups. Initially a method was developed to calculate the area of triangles between three atoms making up the fragment. This was employed systematically as a means of reducing the total number of variables required to describe the geometry of the fragments.

Multivariate statistical methods were also applied with the aim of reducing the number of variables required to describe the geometry of the fragment in a systematic manner. The methods employed were factor analysis and sparse principal components analysis. Both of these methods were used to extract important variables from the original default geometric definition, total geometries. The extracted variables were then used as input for *dSNAP* and were compared with the original output.

Biplots were used to visualise the variables describing the fragments [28, 25]. Biplots are multivariate analogues to scatter plots and are used to visualise how the fragments are related to the variables describing them. Owing to the large number of variables that make up the definition factor analysis was applied to extract the important variables before the biplot was calculated. The biplots give an overview of the correlation matrix and using these plots it is possible to select variables that are influencing the formation of clusters in *dSNAP* .

Declaration

The thesis has been written in accordance with the University and all work presented is original and performed by the author unless otherwise stated and referenced in the text.

Duncan JM Sneddon

Acknowledgements

Firstly I would like to thank my supervisors Prof. Chris Gilmore and Prof. Chick Wilson for giving me the opportunity to get into a very different field of science and for supporting me throughout my research. I would like to recognise the opportunity that I was given to expand my horizons into avenues of science that I would not have ordinarily found myself in. I am also grateful for the continuous encouragement and feedback that I have been given over the past years in their own distinct ways.

I would also like to thank the late Dr. Andrew Parkin who sadly passed away during my research for his invaluable and insightful contribution to my research.

The contribution of the remainder of my peers during has to be recognised: Dr. Gordon Barr for maintaining the program *dSNAP*. Dr. Anna Collins for her invaluable enthusiasm and experience especially during the initial part of my research when I was finding my feet. Dr Lynne Thomas who despite suffering from chronic Welsh-itus, answered endless questions and offered good advice even when it wasn't asked for. Dr. Sooz Buttar and Dr. Marc Schmidtman for being generally and endlessly helpful and encouraging during the whole of my research. Dr Martin Adam for nothing in particular but he'd feel left out if he wasn't mentioned.

I also have to give a general, but not insignificant, thanks to the Gilmore group and the Chicklets at large without whom my research experience would have been boring but much quieter.

I also would like to mention the members Young Crystallographers who have provide such a vibrant and interesting field of research to work in. I would also like to recognise the contribution my friends, both in the Chemistry Department and in the wider world, have offers such fantastic encouragement and by providing many great memories during my time at University and hopefully for many years to come.

I would also like to mention my family and especially my Mother Joan and Father Alasdair for their contribution to my education, unwavering support thought my life, and my upbringing that has given me an open mind that I use as I please. I, after all, wouldn't be here at all if it wasn't for you.

Contents

1	Introduction	13
1.1	Cluster Analysis	14
1.2	The <i>d</i> SNAP program and the clustering methodology.	15
1.3	Fragment Viewer	23
1.4	Variable Space	25
1.5	Wider applications of cluster analysis in molecular sciences	28
1.5.1	Bayesian Methods	30
1.6	Previous Work	31
2	Assessing Total Geometries	33
2.1	Introduction	33
2.1.1	Nature of the variables	35
2.2	Model Examples	35
2.2.1	3-chlorobut-2-ene-thiolate	35
2.2.1.1	Total Geometries	39
2.2.1.2	Clustering with angles only	41
2.2.1.3	Clustering with distances only	42
2.2.1.4	Bonded distances and angles only	43
2.2.1.5	Bonded distances, angles and backbone torsion angle	45
2.2.2	3-aminobutan-2-ol	45
2.2.2.1	Total Geometries	46
2.2.2.2	Angles only	52
2.2.2.3	Distances only	53

2.2.2.4	Bonded distances and angles	54
2.2.2.5	Bonded distances, angles and backbone torsion	55
2.2.3	Pentan-2-one	56
2.2.3.1	Total Geometries	57
2.2.3.2	Angles only	60
2.2.3.3	Distances Only	62
2.2.3.4	Bonded Variables	65
2.2.3.5	Bonded variables and torsion	66
2.3	Conclusions	66
3	Triangles	68
3.1	Introduction	68
3.1.1	Different types of variables	68
3.1.2	Semibonded angles	71
3.2	Triangles as a means of reducing variables	71
3.2.1	Calculating the area of triangles	71
3.2.2	Area of triangles	74
3.2.3	Two dihedral angles with five atoms	75
3.2.4	Modifying coordinates to simulate a torsion	76
3.2.5	Use as input to <i>dSNAP</i>	79
3.3	Conclusions	81
4	Factor Analysis	85
4.1	Introduction	85
4.2	Example: Finding Common Factors Affecting Exam Grades .	90
4.3	Application to <i>dSNAP</i>	91
4.4	3-chlorobut-2-ene-thiolate	91
4.5	3-aminobutan-2-ol	93
4.6	Pentan-2-one	93
4.7	Conclusions	98
5	Biplots	102
5.1	Calculating biplots	104
5.2	Model Examples	108

5.2.1	Difluoroalkene	108
5.3	Reducing Variables	109
5.4	Model examples with reduced variables	109
5.4.1	3-chlorobut-2-ene-thiolate	109
5.4.2	3-aminobutan-2-ol	111
5.4.3	Pentan-2-one	113
5.5	Conclusions	115
6	Sparse principal components analysis	118
6.1	Method	122
6.2	Model examples	122
6.2.1	Difluoroalkene	122
6.2.2	3-aminobutan-2-ol	124
6.2.3	Pentan-2-one	127
6.3	Conclusions	132
7	Conclusions	133
7.0.1	Future Work	136
A	Geometric analysis	138
B	Triangles	141
C	Factor Analysis	147
C.1	3-chlorobut-2-ene-thiolate	147
C.2	3-aminobutan-2-ol	152
C.3	Pentanone-2-one	157
D	Sparse principal components analysis	162
D.1	Difluoroalkene	162
D.2	3-aminobutan-2-ol	164
D.3	Pentan-2-one	166

List of Figures

1.1	An example of a MMDS plot from <i>d</i> SNAP	17
1.2	Example of silhouettes as output by <i>d</i> SNAP.	19
1.3	Scree plot as output from <i>d</i> SNAP.	20
1.4	An example of a dendrogram generated by <i>d</i> SNAP	21
1.5	An illustration of how fragments are joined in the dendrogram during cluster analysis with an accompanying 2 dimensional MMDS plot.	24
1.6	Output from the fragment viewer.	26
1.7	A scatterplot chosen to illustrate that a high correlation does not necessary give a reason for the formation of clusters. . . .	27
2.1	A figure demonstrating how two different variables can be correlated in a data set.	37
2.2	Figures indicating the differences in conformation in the fragment 3-chlorobut-2-ene-thiolate where the geometry of the fragment was defined by total geometries.	40
2.3	The overlay of the fragments 3-chlorobut-2-ene-thiolate	41
2.4	Dendrogram and MMDS plot of the fragment 3-chlorobut-2-ene-thiolate where the geometry of the fragments have been defined by all angles only	43
2.5	Dendrogram and MMDS plot clustered using all distances only for the fragment 3-chlorobut-2-ene-thiolate.	44
2.6	Dendrogram and MMDS plot clustered using bonded angles and bonded distances only for the fragment 3-chlorobut-2-ene-thiolate.	45

2.7	Dendrogram and MMDS plot clustered using bonded angles, bonded distances and carbon backbone torsion angle for the fragment 3-chlorobut-2-ene-thiolate.	46
2.8	Fragment overlay of 3-chlorobut-2-ene-thiolate where the fragment was defined by bonded angles, bonded distances and backbone torsion.	47
2.9	An example of Newman projections	47
2.10	Predicted conformations of the fragment 3-aminobutan-2-ol	48
2.11	Dendrogram and Newman projections of the fragment 3-aminobutan-2-ol. The geometry of the fragments in this dendrogram have been defined by total geometries. The Newman projections represent the conformation of the fragment within that cluster.	50
2.12	MMDS plot clustered using total geometries for the fragment 3-aminobutan-ol.	51
2.13	Dendrogram and MMDS plot clustered using all angles only for 3aminobutan-2-ol.	51
2.14	Dendrogram and MMDS plot clustered using all distances only for the fragment 3-aminobutan-2-ol.	53
2.15	Dendrogram and MMDS plot clustered using bonded angles and distances only for the fragment 3-aminobutan-2-ol.	55
2.16	Dendrogram and MMDS plot clustered using bonded angles, bonded distances and backbone torsion for the fragment 3-aminobutan-2-ol.	55
2.17	An overlay of the fragment 3-aminobutan-2-ol where the fragment was defined by bonded variables and backbone torsion.	56
2.18	Diagram of the fragment pentan-2-one with the free torsions indicated by the blue arrows.	57
2.19	Dendrogram clustered using total geometries for the fragment pentan-2-one.	58
2.20	MMDS plot clustered using total geometries for the fragment pentan-2-one.	59
2.21	Overlay of the fragment pentan-2-one where the geometry was defined by total geometries.	60

2.22	Dendrogram and MMDS plot clustered using all angles only for the fragment pentan-2-one.	61
2.23	Overlay of the fragment viewer 3-aminobutan-2-ol when the fragment was defined by angles only.	61
2.24	Dendrogram and MMDS plot clustered using all distances only of the fragment pentan-2-one.	63
2.25	Fragment view of fragment pentan-2-one when the geometry of the fragment was defined by atomic distances only	63
2.26	Dendrogram and MMDS plot clustered using bonded variables only for the fragment pentan-2-one.	65
2.27	Fragment view of cluster A where the fragment was defined by all bonded variables only.	66
3.1	Sketch of the geometry of a hypothetical molecule	72
3.2	Diagram of simulated dihedral angle.	73
3.3	Graph of triangle area calculated for a range of torsion angles and a single bonded angle.	75
3.4	An illustration of the two torsion angles that are calculated with the triangle indicated.	76
3.5	Description of the program that simulates a five atom fragment being rotated around two torsion angles.	79
3.6	A surface plot representing the area of a triangle between two torsion angles	80
3.7	An illustration of the triangles used to describe the geometry of the fragment pentan-2-one	82
3.8	Combined MMDS plot, cell display and dendrogram of the fragment pentan-2-one.	83
3.9	Fragment view of pentan-2-one where the fragment was defined with triangles	84
4.1	Cell displays of 3-chlorobut-2-ene-thiolate with the fragments defined with total geometries on the top and with the variables reduced by the application of factor analysis on the bottom.	94

4.2	Dendrogram of 3-chlorobut-2-ene-thiolate with the fragments defined with total geometries on the top and with the variables reduced by the application of factor analysis on the bottom.	95
4.3	Cell display of the 3-aminobutan-2-ol with the geometry of the fragments defined by total geometries on the top and with the variables reduced by the application of factor analysis on the bottom.	96
4.4	Dendrograms of the 3-aminobutan-2-ol with the geometry of the fragments defined by total geometries on the top and with the variables reduced by the application of factor analysis on the bottom.	97
4.5	Cell display of the pentan-2-one with the geometry of the fragments defined by total geometries on the top and with the variables reduced by the application of factor analysis (bottom).	99
4.6	Dendrograms of the pentan-2-one with the geometry of the fragments defined by total geometries on the left and with the variables reduced by the application of factor analysis (bottom).	100
5.1	An illustrative biplot with the corresponding correlation matrix and an indication of the standard deviation.	103
5.2	A figure illustrating the properties of the process of interpolation in the context of biplots.	104
5.3	An example of a data set to illustrate the properties of a biplot.	105
5.4	Search information for the fragment difluoroalkene	107
5.5	Dendrogram of difluoroalkene with the fragments illustrated below	109
5.6	Biplot of difluoroalkene with default data matrix	110
5.7	Biplot of the fragment 3-chlorobut-2-ene-thiolate where factor analysis was applied to reduce the number of variables required to describe the fragment.	112
5.8	Diagram of the fragment 3-aminobutan-2-ol	113
5.9	Biplot of the fragment 3-aminobutan-2-ol where the number of variables has been reduced by the application of factor analysis.	114

5.10	Biplot of the fragment pentan-2-one where factor analysis has been used to reduce the number of variables before analysis.	116
6.1	Dendrogram of difluoroalkene with diagram of the fragment.	123
6.2	Cell displays of the fragment difluoroalkene with the fragment described by total geometries on the top and with the variables reduced by the application of sparse principal components analysis on the bottom.	125
6.3	Dendrograms of the fragment difluoroalkene with the fragment described by total geometries on the top and with the variables reduced by the application of sparse principal components analysis on the bottom.	126
6.4	Expected geometric changes in the fragment 3-aminobutan-2-ol. Also there are 2 chiral centres which should also be found in S* and R* con formation.	127
6.5	Cell display comparing the fragment 3-aminobutan-2-ol described with total geometries and with the variables reduced by the application of sparse PCA.	128
6.6	Dendrogram of comparing the fragment 3-aminobutan-2-ol when the fragment was described by total geometries on the top and with the variables reduced by the application of sparse PCA on the bottom.	129
6.7	Cell display comparing the fragment pentan-2-one where the fragment was defined by total geometries (top) and when the number of variables are reduced by the application of sparse principal components analysis (bottom)	130
6.8	Dendrogram comparing the fragment pentan-2-one where the fragment was defined by total geometries (top) and when the number of variables are reduced by the application of sparse principal components analysis (bottom)	131

List of Tables

1.1	Description of the clustering methods available in <i>dSNAP</i> . [23]	22
1.2	Parameters for Equation describing how clusters are joined taken from Everitt <i>et al</i>	23
2.1	Model examples of fragments that will be used thought this thesis.	36
2.2	Table of occurrences of the fragment: 3-chlorobut-2-ene-thiolate	38
2.3	Cluster equivalents; Angles Vs Angles and Distances	52
2.4	Cluster Equivalents: Total geometries against distance	54
3.1	A selection of variables that describe the conformation of the fragment 3-aminobutan-2-ol	70
4.1	Table of factor loadings for the model example	91
6.1	Settings for sparse PCA. algo controls the method for computing the matrix exponential. Gapchange is the target reduction in duality gap. Maxiter in the maximum number of iterations and ρ is a parameter controlling sparsity. Info controls the verbosity of the reporting.	122
A.1	Table that describes whither the fragment pentan-2-one has fallen into the same cluster as when the fragment was described by total geometries.	138
B.1	Table describing the variables of 3-aminobutan-2-ol.	141

C.1	Table of rotated factor loadings of the fragment 3-chlorobut-2-ene-thiolate	147
C.2	Table of commonalities of the fragment 3-chlorobut-2-ene-thiolate	149
C.3	Table of variances of factor analysis of the fragment 3-chlorobut-2-ene-thiolate	151
C.4	Table of rotated factor loadings of the fragment 3-aminobutan-2-ol	152
C.5	Table of commonalities of the fragment 3-aminobutan-2-ol . . .	153
C.6	Table of variances of factor analysis of the fragment 3-aminobutan-2-ol	156
C.7	Table of rotated factor loadings of the fragment pentanone-2-one	157
C.8	Table of commonalities of the fragment pentanone-2-one . . .	158
C.9	Table of variances of factor analysis of the fragment pentanone	161
D.1	First Eigenvalue of DSPCA of the fragment difluoroalkene. . .	162
D.2	First Eigenvector of DSPCA of the fragment 3aminobutan-2-ol.	164
D.3	First Eigenvector of DSPCA of the fragment pentan-2-one. . .	166

Chapter 1

Introduction

“An intelligent being cannot treat every object it sees as a unique entity unlike anything else in the universe. It has to put objects in categories so that it may apply its hard-won knowledge about similar objects encountered in the past, to the object in hand”

Steven Pinker, How the Mind Works.1997

The ability to quickly and accurately interpret structural data mined from the immense numbers of structures currently held within the Cambridge Structural Database (CSD) [1] is a huge asset to structural chemists. As of 1st January 2009 there are more than 469,611 crystal structures available in the CSD [22]. *d*SNAP is a program developed at the University of Glasgow [10]. This program applies cluster analysis and other statistical analyses to the information extracted from the CSD. This program sets out to group specific parts of crystal structures mined from the CSD into groups that are of similar in conformation. This aim of this research is to investigate methods that could be applied to the geometric description of the fragment to reduce the number of variables required to achieve this.

1.1 Cluster Analysis

Classification of objects into groups according to their properties has been ongoing in science for centuries. The work of Aristotle (384 BC – 322 BC), Theophrastos (372 BC – 287 BC) and Linnaeus (1707–1788) underpinned for centuries the classification of plants and animals. The work of Mendeleev creating the first version of the periodic table of the elements is an early example of classification in chemistry. On a very basic level the classification of large datasets into groups that share common features will allow quicker more accurate evaluation of the data. Moreover it will remove some of the human error from the process of the interpretation of large volumes of data while simultaneously uncovering subtle difference or similarities that may have been overlooked.

Given the relative complexity of the geometry of molecular fragments mined from the CSD combined with the potential volume of data, it is necessary to use statistical methods to group fragments into clusters of similar conformations. For this process to progress manually, a method such as binning the fragments into groups according to the knowledge of the investigator could be employed. This would be an extremely long and tedious process that is fraught with pitfalls, not least of which is that the binning could be based on assumption not on observation. The basic premise of structural prediction and crystallography is that conformations that are found in the crystalline environment are assumed to be of a low energy conformation [5]. By examining a portion of the molecule that was originally crystallised, there is a possibility that the conformation of this portion of a molecule will be affected by the chemical context from the original molecule. These differences, if present, should give rise to populations of fragments with different conformations. These are the differences that give rise to the formation of clusters in *d*SNAP.

*d*SNAP is a program that is used in conjunction with the CSD. Initially the CSD is queried using the program ConQuest[11] created by the Cambridge Crystallographic Data Centre (CCDC). Typically a portion of a molecule will be drawn and searched for within conquest. This portion of a

molecule will be known as the “fragment”. This search will produce a number of fragments which in this research will be termed “hits” and there may be many hits within a single molecule of structure in the CSD. The input to *dSNAP* is the coordinates of the atoms for each hit in the search. These coordinates are then processed within *dSNAP* to produce the definition of the geometry of the fragment. The default definition for the fragment is termed ‘total geometries’. This geometric is made up of all of the distances and all of the angles between all of the atoms in a fragment including the non-bonded interactions. A discussion of the merits of this particular geometric definition takes place in the following chapter. The description of the geometry is now represented as a list of positive scalar values. These values are then used to carry out the calculations within *dSNAP*.

1.2 The *dSNAP* program and the clustering methodology.

The geometric data is represented as a matrix with n hits (samples) represented by p variables. The geometric information mined from the CSD is converted into a symmetric ($n \times n$) Minkowski distance matrix, \mathbf{d}^s using the following formula:

$$d_{ij}^s = \left(\sum_{k=1}^m w_k |x_{ik} - x_{jk}|^\lambda \right)^{\frac{1}{\lambda}} \quad (1.1)$$

where x_{ik} and x_{jk} are the k th variables of the i th and j th sample respectively and w_k is a weighting that can be applied to each of the variables. In *dSNAP* this is set to one by default. λ is a user selectable parameter in *dSNAP*, the default value is two which corresponds to a Euclidean distance matrix. A value of one can be used if a city block distance matrix is desired. The variables, x are distances and angles between the atoms of the fragment mined from the CSD. The superscript ‘s’ for the matrix \mathbf{d} indicates the matrix is in subject or stimulus space in order to distinguish it from the related variable space. The distance matrix, calculated in Equation 1.1,

is then standardised by dividing each variable by its sample range to give $0 \leq d_{ij}^s \leq 1.0$ and $d_{ij}^s = 1.0$. [10] This is done so that each of the variables make equal contributions. If standardisation was not carried out, the atomic distances would not be fairly measured as these variables will have much lower variance than the atomic angles.

Metric multidimensional scaling (MMDS) [18] is used to generate a three-dimensional Euclidean space in which each of the fragments is represented as a single point within this space. A simple definition of multidimensional scaling is a search for a lower dimensional space, usually Euclidean, where each of the points in the space represents a single fragment. The points are placed in such a way that the distances between points in the lower dimensional space are placed to approximate the distances calculated using Equation 1.1. Using the distance matrix \mathbf{d}^s , a matrix $\mathbf{A}_{(n \times n)}$ is constructed.

$$A = -\frac{1}{2} \left(I_n - \frac{1}{n} \mathbf{i}_n \mathbf{i}_n' \right) D^s \left(I_n - \frac{1}{n} \mathbf{i}_n \mathbf{i}_n' \right) \quad (1.2)$$

Where \mathbf{I}_n is an $(n \times n)$ identity matrix, \mathbf{i}_n is an $(n \times 1)$ vector of unities and \mathbf{D}_s is a matrix of squared distances. The eigenvectors of \mathbf{A} , $\nu_1, \nu_2 \dots \nu_n$ form a vector \mathbf{V} and the corresponding eigenvalues $\lambda_1, \lambda_2 \dots \lambda_n$ give a matrix Λ . A total of p eigenvalues are selected to be positive and the remaining $(n - p)$ eigenvalues are set to zero. A set of coordinates in p dimensions can be defined via the matrix $\mathbf{X}_{(n \times p)}$

$$\mathbf{X} = \mathbf{V} \Lambda^{\frac{1}{2}} \quad (1.3)$$

in *d*SNAP p is set to three to give three dimensions and the matrix \mathbf{X} can be used to plot each of the fragments mined from the CSD into the three-dimensional Euclidian space [10, 18]. An example of the MMDS plot generated by *d*SNAP can be seen in Figure 1.1.

For each of the clusters with three or more members, a most representative sample (MRS) is highlighted in the MMDS plot. The MRS is defined as the member of a cluster that has the minimum distance to all of the other members of the cluster. For example, for cluster J containing m patterns,

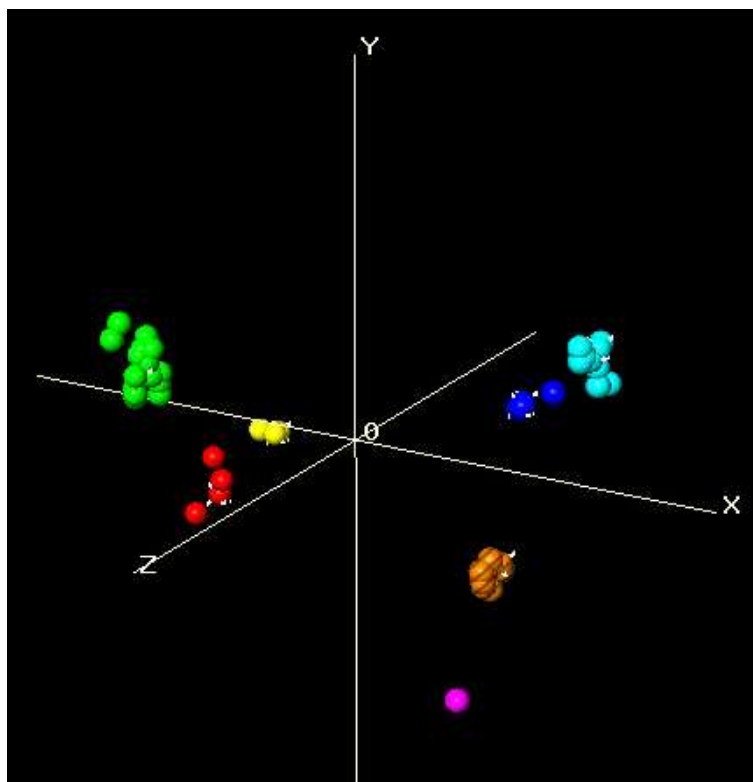


Figure 1.1: An example of a MMDS plot from d SNAP. The distance matrix which in this case is a 51×51 matrix has been reduced to a 51×3 matrix using MMDS above. The points that represent the 51 samples have been plotted into the 3 dimensional space such that the distances in the lower dimensional space are fitted in such a manner as to approximate the distances calculated using Equation 1.1. As a result of this, the proximity of the samples, or in this case, fragments are an indication of their similarity. The closer the points are in space the more similar the fragments are in conformation. The most representative samples are marked with a white cross.

the most representative sample, i , is defined

$$\min \left[\sum_{\substack{j=1 \\ i,j \in J}}^m d(i,j)/m \right] \quad (1.4)$$

Checks in the MMDS calculation are carried out to ensure that the data can be reduced to three dimensions without losing the essential features of the data. The first of these is the generation of a distance matrix from the $\mathbf{X}_{(n \times 3)}$ and the element-by-element comparison with the original distance matrix \mathbf{d}^s using a mean of the Pearson [49] and Spearman correlation coefficients [53].

Scree plots and silhouettes [51, 9] are employed to validate the clusters. Silhouettes are calculated by firstly calculating the dissimilarity coefficient δ_{ij} .

$$\delta_{ij} = d_{ij}/d_{ij}^{\max} \quad (1.5)$$

If the fragment i belongs to cluster C_r which contains n_r structures,

$$a_i = \sum_{\substack{j \in C_r \\ j \neq i}} \delta_{ij}/(n-1) \quad (1.6)$$

and

$$b_i = \min_{s \neq r} \left(\sum_{j \in C_s} \delta_{ij}/n_s \right) \quad (1.7)$$

The silhouette, h_i , for fragment i is then

$$h_i = \frac{b_i - a_i}{\max(a_i, b_i)} \quad (1.8)$$

Silhouette values are assigned to all members of a cluster and give an estimate of the membership for each fragment to that cluster. h_i lies between -1.0 and 1.0 and the results are plotted on a histogram, this allow clear identification of outliers in a cluster. Each cluster should have a tight silhouette with few or no outliers and all the values should be greater than 0. Ideal

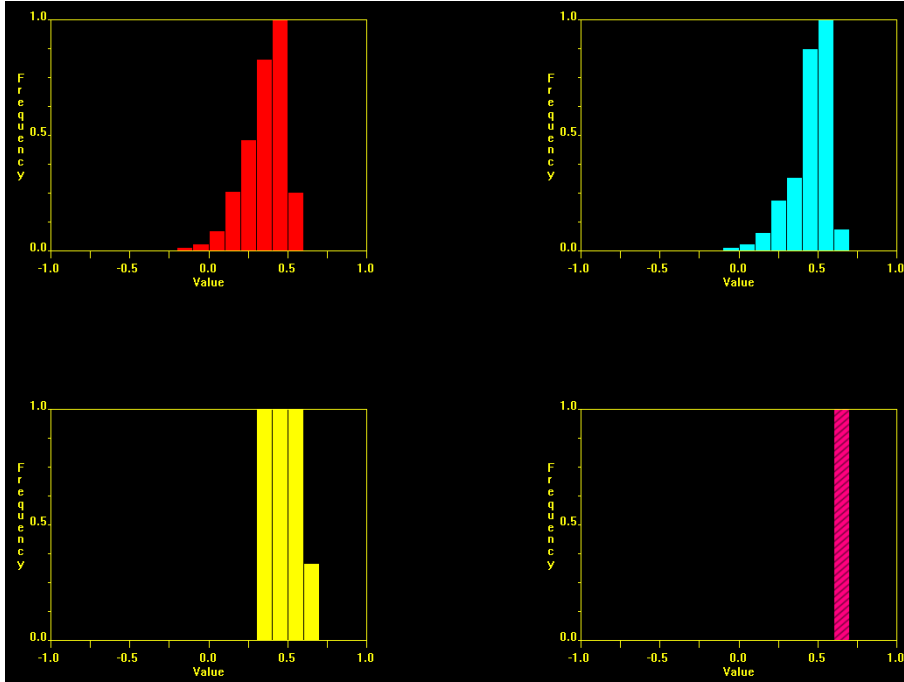


Figure 1.2: Example of silhouettes as output by d SNAP. In this example the silhouettes are well defined and there are no outliers.

clusters have a $h_i \geq 0.5$.

The scree plot is also used to validate the clustering process. Using principal component analysis of the matrix \mathbf{A} a set of sorted eigenvalues are produced and plotted in a scree plot. The scree plot should have a steep decent with no dramatic changes in gradient. Both of these tools are used to check the quality of the clusters and the input data.

d SNAP employs a clustering algorithm similar to that described in [26, 8, 7, 9]. The clustering portion of this algorithm is based on hierarchical cluster analysis. Hierarchical cluster analysis begins with each of the fragments mined from the CSD search as single cluster with a single member. That is, initially there will be n clusters made up of one fragment. Upon the completion of this process there will be a single cluster containing n fragments.

At the beginning of the clustering process the fragments closest together when defined by the distance matrix calculated using Equation 1.1 are joined

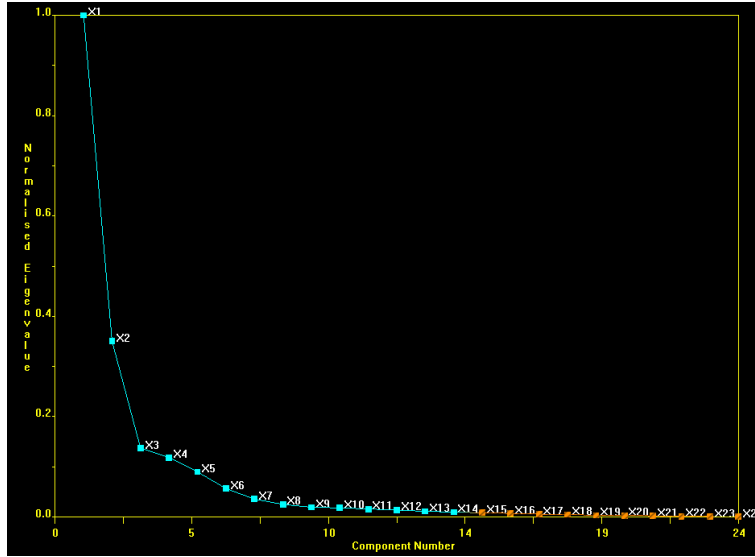


Figure 1.3: Scree plot as output from *d*SNAP. This is a good example of a scree plot, there are no dramatic changes in gradient and the gradient falls quickly. The change in colour at X14 indicates the 14 components can explain 95% of the data.

and are regarded are now regarded as a single cluster [44]. Now that two fragments have been joined into a single cluster there is a problem of defining how to define the distance between the new cluster and any of the other fragments or clusters. When two classes or clusters (C_i and C_j) are joined there is a problem of defining the distance between the newly formed class $C_i \cup C_j$ and the other classes C_k . There are a number of different ways of doing this but the methods employed in *d*SNAP are described in Table 1.1 and the α , β and γ terms are defined in Table 1.2. The distance between the new class formed by merging C_i and C_j and any other class C_k is given by Equation 1.9.

$$d(C_i \cup C_j, C_k) = \alpha_i d(C_i, C_k) + \alpha_j d(C_j, C_k) + \beta d(C_i, C_j) + \gamma |d(C_i, C_k) - d(C_j, C_k)| \quad (1.9)$$

where d is the distance between the new class and any other class.

On the completion of cluster analysis, a dendrogram is drawn. A dendro-

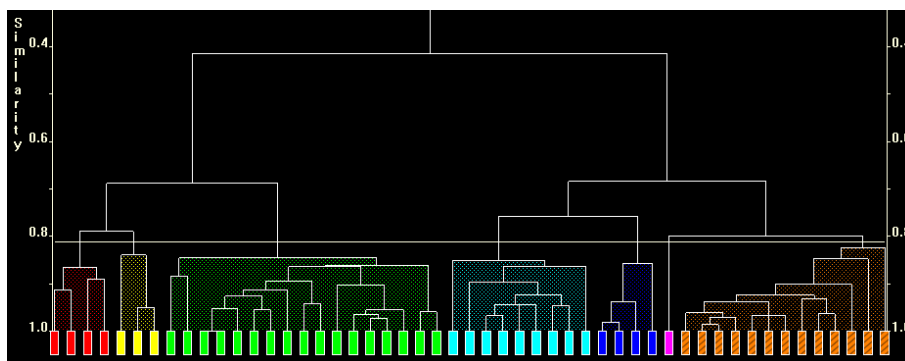


Figure 1.4: An example of a dendrogram generated by *dSNAP*. The coloured blocks at the bottom of the plot are the individual samples or fragments in this case. The y axes is an arbitrary measure of similarity that is specific to each analysis. In order to give an indication of the similarity between the fragments the fragments or groups of fragments are joined together using horizontal lines known a tie bars. The lower on the y axis the tie bar is the greater the level of similarity between these fragments.

gram is a tree like diagram where each of the fragments are represented as individual ‘leaves’ at the bottom of the diagram, in the case of the dendrogram drawn in *dSNAP*. The program gives the five options for the generation of dendrograms. These are: single link, complete link, weighted average link, centroid and group average link. The group average link method is the default option in *dSNAP*[10]. Table 1.1 shows the criteria by which the clusters are joined along with remarks about the formation of clusters from Everitt *et al* [23]. The fragments are then joined together in the dendrogram using horizontal lines that indicate how similar these clusters are. In the case of *dSNAP* the y axis illustrates at what level of similarity these clusters are joined. This similarity is an indication of the distances between clusters calculated using Equation 1.1 and should not be regarded as an absolute measure of similarity between fragments mined from the CDS.

The number of clusters is defined by the cut level, represented by a moveable horizontal bar on the dendrogram. The initial position of this horizontal bar is estimated using two methods, the first eigenanalysis carried out using the \mathbf{A} matrix from the MMDS calculation (Equation (1.2)) and the other is eigenanalysis of the correlation matrix ρ , corresponding to d :

Method	Distance between cluster defined as:	Remarks
Single link method	Minimum distance between pair of objects, one in one cluster, one in the other	Tends to produce unbalanced and straggly clusters ('chaining'), especially in large data sets. Does not take account of cluster structure.
Complete link method	Maximum distance between pair of objects, one in one cluster, one in the other	Tends to find compact clusters with equal diameters (maximum distance between objects). Does not take account of cluster structure.
Weighted average link	Squared Euclidean distance between weighted centroids	Assumes points can be represented in Euclidean space for geometrical interpretation. New group intermediate in position between merged groups, Subject to reversal.
Centroid	Squared Euclidean distance between mean vectors (centroids)	Assumes points can be represented in Euclidean space (for geometrical interpretation). The more numerous of the two groups clustered dominates the merged clusters, subject to reversals.
Group average link	Average distance between pair of objects, one in one cluster, one in the other	Tends to join clusters with small variance. Intermediate between single and complete linkage. Takes account of cluster structure. Relatively robust.

Table 1.1: Description of the clustering methods available in *d*SNAP. [23]

$$\rho = 2\mathbf{d}^s - \mathbf{I} \quad (1.10)$$

where \mathbf{I} is the identity matrix and \mathbf{d}^s is the distance matrix calculated using Equation 1.1. In both cases the eigenvalues of the relevant matrix are sorted in descending order. Once 95% of the variability is accounted for, the number of eigenvalues is selected and this is used to define the number of clusters. Since two methods are used, the results are averaged. There is no formally correct mathematical method to calculate the position of the cut level in cluster analysis; as a result the position of the cut level is only an estimate but the user interface used by *d*SNAP allows the user to change the cut level easily and accurately to best represent the underlying chemistry of the fragment being investigated [10].

The tools used in *d*SNAP most frequently are the MMDS plot (Figure 1.1) and the dendrogram (Figure 1.4). Both of these plots show a representation of the distance matrix. In the MMDS plot, these differences are illustrated as the difference in distance between the spheres in the plot. Spheres in close

Method	α_i	β	γ
Single linkage	$\frac{1}{2}$	0	$-\frac{1}{2}$
Complete linkage	$\frac{1}{2}$	0	$\frac{1}{2}$
Average link	$n_i(n_i + n_j)$	0	0
Weighted average link	$\frac{1}{2}$	0	0
Centroid	$n_i/(n_i + n_j)$	$-n_i n_j / (n_i + n_j)^2$	0
Sum of squares	$(n_i + n_k)/(n_i + n_j + n_k)$	$-n_k/(n_i + n_j + n_k)$	0

Table 1.2: Parameters for Equation 1.9 taken from Everitt *et al* [23]

proximity represent fragments that are closely related in structure. Much in the same way dendrograms illustrate the similarity between fragments or indeed clusters by the height of the tie bar joining them together. A lower tie bar indicates a greater degree of similarity. These tools are used in unison to explore the relationship between the structures of the fragments. The 3D plot generated by metric multidimensional scaling gives a representation of the structure of the data, such as clusters that are merging or are actually continuous. This information is also displayed in the dendrogram but can be difficult to spot. The advantage that the dendrogram has is that a high dimensional dataset is displayed in a two dimensional manner and as such is more applicable to being used as illustrations. This is illustrated in Figure 1.5.

Once the relationship between fragments has been established, it is then necessary to examine the fragments and the variables describing the fragments that have formed the clusters, to understand what conformation the fragments are in.

1.3 Fragment Viewer

There is also a viewer that allows individual fragments to be overlaid. In this context there are n fragments with precisely the same number of points. This situation lends itself well to Procrustes analysis[18, 27, 34]. Procrustes analysis is a process where the coordinates of, in this case a molecular fragment, are rotated, translated, reflect and dilated in such a way that the fragment's coordinates are minimised in a least squared sense with another

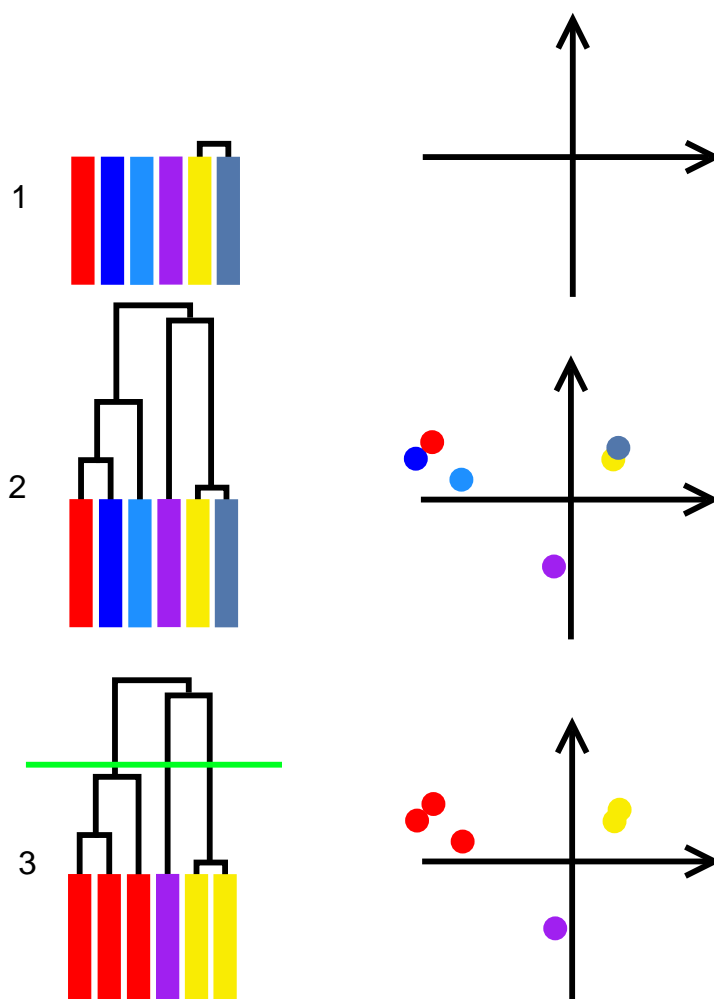


Figure 1.5: An illustration of how fragments are joined in the dendrogram during cluster analysis with an accompanying 2 dimensional MMDS plot. Initially all of the fragments are regarded as being unique and unrelated. As cluster analysis progresses the fragments that are closest together according to the distance matrix calculated using Equation 1.1 are joined. This is represented as point one in the above figure. At point two all of the fragments have been joined together according to their distance and therefore their similarity. In the dendrograms generated by *dSNAP* the lower the tie bar the closer the fragments are in distance and therefore similarity. By examining the MMDS plot to the right of the figure the relationship between the height of the tie bars and the distance between points representing samples is illustrated. Finally, at point three a cut level was applied to the dendrogram where all of the cluster of fragments below this level are regarded as being a single group and are coloured accordingly.

fragment. Let \mathbf{X} be a matrix represent the coordinates of the first fragment and the matrix \mathbf{Y} represent the coordinates of the second fragment. The sum of squared distance between the points is given below

$$R^2 = \sum_n^{r=1} (\mathbf{y}_r - \mathbf{x}_r)^T (\mathbf{y}_r - \mathbf{x}_r) \quad (1.11)$$

where $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n]^T$ and $\mathbf{Y} = [\mathbf{y}_1, \dots, \mathbf{y}_n]^T$ and \mathbf{x}_r and \mathbf{y}_r are the coordinate vectors of the r th point within each of the fragments.

The vector coordinates \mathbf{x}_r are modified using the equation below to produce the vector coordinates \mathbf{x}'_r .

$$\mathbf{x}'_r = \rho \mathbf{A}^T \mathbf{x}_r + \mathbf{b} \quad (1.12)$$

where matrix \mathbf{A} is orthogonal giving a rotation and potentially a reflection, vector \mathbf{b} is a rigid translation vector and ρ is the dilation. The application of this formula seeks to minimise the new sum of squared distances between points

$$R^2 = \sum_n^{r=1} (\mathbf{y}_r - \rho \mathbf{A}^T \mathbf{x}_r - \mathbf{b})^T (\mathbf{y}_r - \rho \mathbf{A}^T \mathbf{x}_r - \mathbf{b}) \quad (1.13)$$

The fragment viewer implemented within *dSNAP* does not use the dilation portion of Procrustes analysis as this may mask any systematic differences in bond length which could be of interest. This is also a method that allows the user to select specific atoms such that Procrustes analysis is only carried out on the selected atoms. This is especially useful when the user would like to emphasise specific differences in conformation. An example of the output from the fragment viewer can be seen in Figure 1.6.

1.4 Variable Space

dSNAP allows the variables describing the fragments to be examined. This feature of *dSNAP* uses what is known as the variables space. The angles and distances, of which there are m variables, are subjected to a Pearson

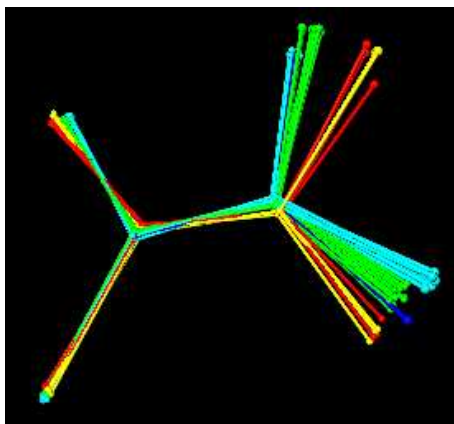


Figure 1.6: Output from the fragment viewer. In this example fragments have been selected from the dendrogram in Figure 1.4 and have been overlaid. Procrustes had been applied to the left hand atoms in the fragments while the remaining atoms are not overlaid. This feature allows specific features to be exaggerated.

correlation coefficient with every other variable in order to generate a $(m \times m)$ correlation matrix. From this correlation matrix, a distance matrix is calculated using Equation 1.10. Using these matrices a dendrogram and MMDS plot are generated in exactly the same manner as when the fragments mined from the CSD are clustered.

In appearance these plots are similar to those generated when fragments are being compared but these results are interpreted differently (Figures 1.4 and 1.1) [10]. The major difference is that instead of clustering the mathematical distances between fragments, the correlation between variables is clustered. In these plots the variables that are closely related have a high correlation coefficient. That is to say, the dendrogram variables which are highly correlated will be joined by lower tie bars and in the MMDS plot the spheres representing the variables will be closer together. Unfortunately, as a rule, there is little relationship between high correlation between variables and an explanation for clustering. This proves to be a problem because the variables that are highly correlated are not necessarily the ones that distinguish what is causing clusters of fragments to form. With reference to Figure 1.7, it is obvious that no clear universal reason for the formation of all of the

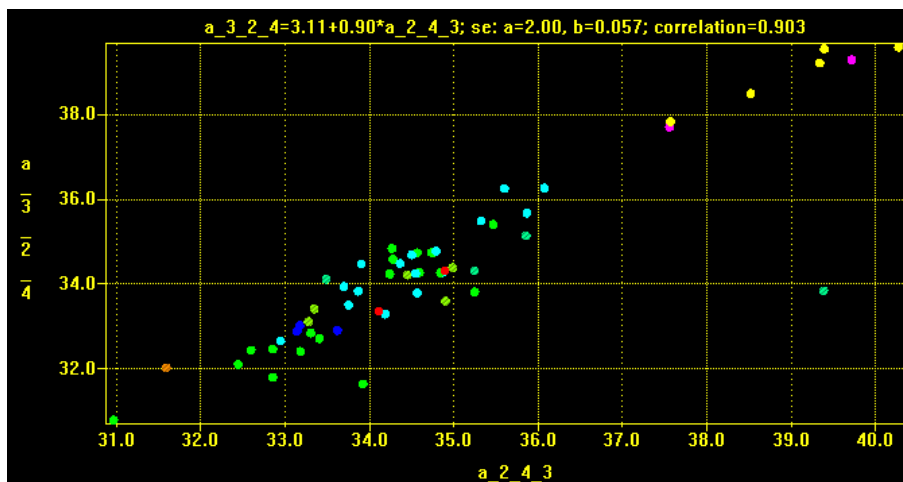


Figure 1.7: A scatterplot chosen to illustrate that a high correlation does not necessary give a reason for the formation of clusters. In this plot the colour of a point was taken from the dendrogram calculated within *d*SNAP. The plot indicates that there is no clear distinction between the clusters indicated by the different colours of the points within this plot.

clusters can be obtained using these variables alone. While it is unreasonable to expect that a single pair of variables will explain the reasons behind the formation of clusters, it is apparent that there is no direct relationship between correlation and useful justification to formation of clusters. In part, this research will aim to identify variables or groups of variables that can be used to justify the formation of clusters.

In spite of this situation, there must be variables that are causing the clusters to form in subjective space. Subjective space is where the fragments mined from the CSD are being compared to one another opposed to variables space where the variables are being compared. Finding variables that can justify the formation of clusters would be useful to analyse the clustering in *d*SNAP. One of the problems is that there are so many variables defined by *d*SNAP. Using the standard definition of total geometries the number of variables is defined as:

$$N_{(bonds)} = \frac{n}{2}(n - 1) \quad (1.14)$$

$$N_{(angles)} = \frac{n}{2}(n - 1)(n - 2) \quad (1.15)$$

where n is the number of atoms. The number of variables increases to the order n^3 with 20 atoms being the technical limitation of *d*SNAP where the geometry of the fragments are described by 3610 variables. The high number of variables describing a fragment of 20 atoms results in approximately 13 million scatterplots, any number of which could yield information critical to the justification of the formation of clusters. It is therefore necessary to have tools that aid the detection of these key variables.

1.5 Wider applications of cluster analysis in molecular sciences

Cluster analysis has many applications in the design of potential drug targets and in molecular biology at large. As a consequence of industrial scale DNA sequencing projects such as the Human Genome Project [17] there is now a huge volume of genomic DNA sequence data across many species. This large volume of data has resulted in the development of many different analytical tools to aid the understanding of these data. One such example is the development of CLUSTAL [33, 39] which is a tool to align sequences of amino acids or nucleotides according to their similarities. Sequences of amino acids or nucleotides are aligned in CLUSTAL using a substitution matrix such as the BLOSUM62 [32] matrix where a score is given for a substitution of different amino acids such that there is a low score for similar amino acids and higher scores for dissimilar amino acids. There are also penalties for introducing and extending gaps in sequences. When combined in the algorithm produces a pair of aligned sequences. These sequences are then placed onto guide trees that are calculated according to the evolutionary distances in the dataset [57]. The ultimate aim is similar to the aims of this research. That

is, to group objects together according to their similarity such that common features can be explored. In the context of sequence alignment common domains in the amino acid sequence of proteins between species can be examined, for instance. The volume of data that can easily be mined from various sources means that sequence alignment tools are essential for the exploration of sequence data.

Cluster analysis also has a useful application in the identification of compounds that may have therapeutic value. Many systematic searches for drug target begin with structural information about the target molecule which will typically be a protein. This information can be found using macromolecular crystallography. Once a target binding site is established the search for a molecule that can interact with this site begins. This process can be greatly aided if there is already a compound of known activity interacting with the target protein.

At this point knowledge of the interactions that characterise a pharmacophore is useful and where study of the interactions that make up a small molecule crystal structure is extremely useful [48]. The use of the information contained in the CSD means that information about molecular interactions such as hydrogen bonds can be searched for and examined in detail. This information can then be applied to designing molecules that may well bind to the target site of a protein molecule.

Pascard [48] gives an overview of the analysis that has been carried out examining interactions which are common in protein-substrate interaction. This was carried out by examining the interactions and plotting histograms of the polar coordinates. These histograms were then used to characterise the interaction. In this group, *d*SNAP has been used to classify a number of intermolecular interactions using total geometries to cluster fragments into groups with common orientation of the interaction [13, 14, 47].

Once a target has been identified there are also other uses of cluster analysis in the context of drug design. One such example is the use of cluster analysis to aid in the interpretation of quantitative structure-activity relationship between a sample population of drug targets. This method involves combining a number of physical measurements of the compounds in question,

such as molecular weight. Cluster analysis can then be used to organise the compounds into groups from which compounds with desirable characteristics can then be extracted [30].

1.5.1 Bayesian Methods

While cluster analysis and the other methods used in *d*SNAP do not require any prior knowledge of the potential conformations of fragment, there are model based statistical methods which can be applied to the problem of classifying molecular fragments. The work of Perez *et al* [50] utilised Bayesian methods to group fragments into groups of similar conformation where the fragment were ring structures and were defined using torsion angles. This work was approached from two directions: Using *a priori* knowledge of the fragment under investigation to classify the fragments and using the frequency of occurrence as well as the standard deviation of the measurement to classify the population of fragments. These approaches were known as the ‘Classification method’ and ‘Full Bayesian analysis method’ respectively. These methods were used to classify a sample of eight membered rings into groups of similar conformation. Initially, *a priori* knowledge was used to construct an ideal range of conformations and the probability of a fragment being in a specific conformation was calculated and used to classify which conformation the fragment belonged to. Interestingly, this method could also be used to classify fragments that did not necessarily fall into a discrete category. The full Bayesian method generates histograms of the number of preferred conformations and then can calculate the probability of a specific fragment being in one or more conformations. This work was extended by Kessler *et al* [38] to examine cyclic copper complexes where Bayesian methods were applied to data mined from the CSD. This classification was preceded by the application of cluster analysis with the aim of easing the task of deciding the number of clusters. This research also used molecular mechanics calculations to understand the interconversion pathways between conformations.

1.6 Previous Work

The body of work that immediately preceded this thesis is the work of Allen and Doyle [4, 2, 3, 6] where fragments that were mined from the CSD were clustered using a variety of clustering algorithms such as single linked, complete linkage and Jarvis-Partick. [23, 35] The geometry of the fragments in this research was defined using torsion angles. The authors also used principal components analysis to aid in the conformational justification for the formation of clusters. The fragments used in this research were six member carbon rings the symmetry of which required the fragment to be renumbered in a consistent manner. The research of Murray-Rust and Raftery [45, 46] used a least squares method where the difference is calculated by the sum of the squared differences between the Cartesian coordinates of the molecules. More recently Weng *et al* [59] used similar methodologies to Murray-Rust and Raftery to cluster fragments into groups. In both of these papers distance matrices were calculated and cluster analysis was applied with the aim of forming groups of fragments with similar conformations. This was achieved by utilising different clustering algorithms to much the same effect as in this research. *dSNAP* primarily differs from these methods by the definition by which the fragments are defined in the program. In previous research by Allen *et al* [2, 3, 6, 4] the fragments were defined by torsion angles. Since torsion angles are circular measures and are signed (+ or -) particular attention to this was required on the part of the researchers. The other method used by Murray-Rust and Raftery [45, 46] and Weng *et al* [59] used the sum of squared difference between atomic positions to describe the conformation of the fragments. This definition gives a good overview of the difference between fragments but using this definition for the geometry of the fragments does not allow investigation into the reasons behind the formation of clusters using the variables. This research aims to use a geometric definition to overcome these problems. This definition must be automatically applied to any fragment without any prior knowledge of the fragment's expected conformation while still allowing the possibility of examining the variables with the intention of understanding the formation of clusters.

In the following chapters there will be an investigation into the different geometric descriptions that can be applied to the fragments mined from the CSD. Following this there is an attempt to reducing the number of variables by describing groups of variables as the area of a triangle described by these variables. The remainder of this thesis will focus on reducing the number of variables by the application of multivariate statistical methods which will attempt to systematically reduce the number of variables necessary to accurately describe the conformation of the fragments. These methods are factor analysis [52] and sparse principal components analysis [21]. There is also a chapter that will examine the application of biplots to *d*SNAP [28, 29]. Biplots are a means to examine the fragments and the variables describing them in a single plot.

Chapter 2

Assessing Total Geometries

2.1 Introduction

*d*SNAP is a program which will organise a series of fragments into groups of fragments which have similar conformations. These fragments are a motif of atoms that are searched for in the Cambridge Structural Database (CSD) [1]. When the motif is found within the database, the coordinates of these motifs are then output and used as the input information for *d*SNAP. This should then give a population of fragments that have been derived from their original molecules such that the fragments will have different conformations or geometries according to the chemical context from which the fragment was derived.

It is therefore critical that the manner in which the geometries of the fragments under investigation are defined must be robust, unbiased and comprehensive. In other words, this definition has to be universally applicable to all fragments regardless of the chemical nature of the fragment. The definition should be fully automatic and with no need for prior knowledge from the user as this may introduce bias which will adversely affect the analysis. It is also essential that the definition of the fragment will consistently allow the clustering algorithm to run while still leaving the burden of interpretation on the user. This is fundamental to the functioning of a program designed for a user with little or no experience in statistics. Currently, the default definition

for fragments is termed ‘total geometries’. This definition is defined as all angles and all distances between the atoms of the fragment. This includes non-bonded angles and distances as well as bonded angles and distances. The most obvious downside of this definition is that the number of variables will increase significantly as the number of atoms increases. The relationship can be explained using the Equation 1.14. The purpose of this chapter is to investigate different geometric definitions of the fragments of molecules with the aim of uncovering the optimum method of describing the geometry of fragments.

Within this section there will be a number of definitions examined and illustrated with examples. The different definitions which are to be explored are:

- All distances and angles (total geometries)
- All angles
- All distances
- Bonded angles and distances
- Bonded angles, distances and torsion angles

Where the object of this exercise is to find the optimum method of describing the geometry of the fragment being investigated.

As described in the previous chapter, *d*SNAP uses cluster analysis and multidimensional scaling to group or cluster fragments which have a similar shape. The results are then displayed in a manner that allows a user to quickly identify these groups and using visualisation tools, assign chemical meaning to these groups. Initially the fragments will be examined using total geometries as a benchmark. The angle and distance components of total geometries are then examined independently. Finally the bonded variables and the bonded variables with backbone torsions are examined. The torsion angles have been added to the bonded variables in order to distinguish between the rotational conformational changes of a fragment. For example, the relative position of functional groups at either end of a single bond will not be detected by bonded variables alone.

2.1.1 Nature of the variables

The atomic nature of the input data combined with the manner in which the geometry of the fragments is derived means that there can be high correlations between variables. An illustration of this point can be seen in Figure 2.1 where two angles are represented in a schematic form in order to demonstrate why the variables can have such a high correlation. There are only two variables in this example but it is possible to predict that most of the variables will take values to reflect this change in conformation. It must be noted that there are also variables that will not be altered in any significant way by the conformational change illustrated in Figure 2.1. This is the basis of the redundancy of the data. It also illustrates why a single variable does not uniquely describe a change in conformation. This is the result of the fact that each atom's position is described by multiple variables within the fragment. Also, a single variable is describing the relative position of at least two atoms. While this is a problem when trying to identify geometric changes of the fragment from changes in variables, cluster analysis is well suited to redundant datasets since the generation of the distance matrix is the result of a comparison between all of the variables describing each fragment with all of the other fragments (Equation 1.1). Since all fragments are compared with all others, the redundancy does not compromise the integrity of the calculation. This means that a robust definition that can accurately describe the geometry of the fragments may be advantageous over one that uniquely describes the geometry of a fragment in this particular context. It could be an advantage to choose a robust definition to describe the geometry of the fragments and accept the increased computational cost of using a redundant definition of the fragments.

2.2 Model Examples

2.2.1 3-chlorobut-2-ene-thiolate

This fragment is a simple example that has discrete conformational changes that can be easily detected. This change is a *cis/trans* change around the

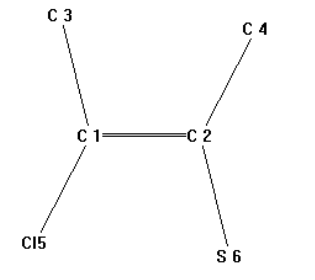
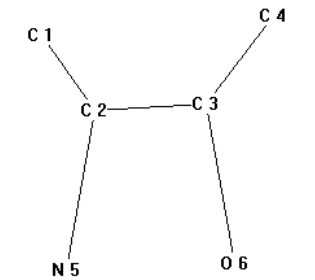
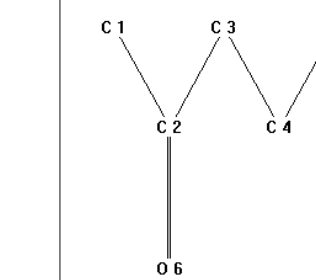
Fragment	3-chlorobut-2-ene-thiolate	3-aminobutan-2-ol	Pentan-2-one
Figure			
CSD version	5.27(November 2005)	5.27(November 2005)	5.29(November 2007)+Updates (Jan 08)
Restrict info	No Refcode restriction	No Refcode restriction	No Refcode restriction
Filters	None	Organics only	3D coordinates determined, Not disordered, Not polymorphic, No powder structures $R \leq 0.05$, No errors No ions, Only Organics
Bond Restrictions	None	None	Bonds 1-2,2-3,3-4,4-5 acyclic
Atom Restrictions	None	Nitrogen restricted to 3 bonded atoms Oxygen restricted to 2 bonded atoms	None
Number of fragments	41	58	113

Table 2.1: Model examples of fragments that will be used thought this thesis.

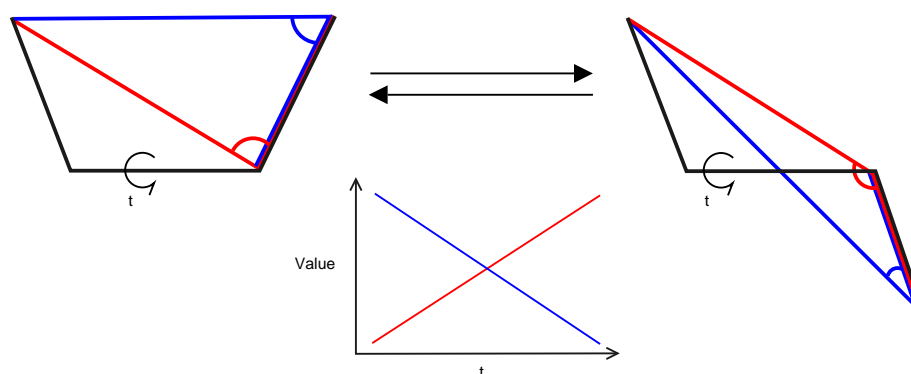


Figure 2.1: A figure demonstrating how two different variables can be correlated in a data set. In this illustration there is a hypothetical fragment described by two related variables. Both of the variables are angles that are measuring the relationship between the bond on the left hand side of the fragment and the atom at the far right. This figure shows how the values of these two variables will change during a continuous rotation around t and that they are negatively correlated.

carbon-carbon double bond. This dataset will provide a straightforward case which will allow different geometric definitions of this fragment to be explored and the results of the cluster analysis to be justified from a chemical perspective. Since this fragment has a double bond at its centre it should have a similar range of conformations to the fragment difluoroalkene. The conformation of this fragment is described in previous work described by this group[10]. In summary, there are two major conformational changes: A *cis/trans* conformation about the double bond, where the carbon atoms are either on the same side as one another or on opposite sides of the double bond; and a restriction in the bonded angles as the chemical context of the fragment changes. For example, if the fragment was derived from a five atom ring, the bonded angles will be smaller where the atoms form part of this ring. The inclusion of a sulfur atom in this fragment means that this atom can bond to other atoms within the molecule which the fragments are derived from. This gives greater conformational possibilities than the difluoroalkene in [10].

Table 2.2 gives a list of occurrences and summary of possible conforma-

Fragment Name	<i>cis/trans</i>	Constrained backbone	Sulfur bonded
AFESIO	<i>cis</i>	No	No
BELWOE	<i>trans</i>	No	No
BELWUK	<i>trans</i>	No	No
COQVIN	<i>cis</i>	Yes	Yes
DATWIF	<i>cis</i>	Yes	Yes
DATWIF_02	<i>cis</i>	Yes	Yes
EDEZAO	<i>trans</i>	Yes	Yes
FAZVOS	<i>cis</i>	Yes	No
FAZVOS_02	<i>cis</i>	Yes	No
FEBSIP	<i>trans</i>	Yes	Yes
FUMBAQ	<i>trans</i>	Yes	Yes
FUMBEQ10	<i>trans</i>	Yes	Yes
GEPVUS	<i>trans</i>	Yes	Yes
GEPVUS_02	<i>trans</i>	Yes	Yes
GILVIH	<i>cis</i>	No	Yes
HAWXUY	<i>trans</i>	Yes	Yes
HAWYAF	<i>trans</i>	Yes	Yes
HOQBEU	<i>cis</i>	Yes	Yes
LAXZAM	<i>cis</i>	Yes	Yes
LAXZEQ	<i>cis</i>	Yes	Yes
LAXZIU	<i>cis</i>	Yes	Yes
LAXZOA	<i>cis</i>	Yes	Yes
LIJDEO	<i>cis</i>	No	No
MAYTOW	<i>cis</i>	Yes	Yes
MAYTOW_02	<i>cis</i>	Yes	Yes
MAYVAC	<i>cis</i>	Yes	Yes
MOSTIX	<i>cis</i>	No	Yes
MOSTIX01	<i>cis</i>	No	Yes
MOSTOD	<i>cis</i>	No	Yes
MOSTUJ	<i>cis</i>	No	Yes
NAWDEV	<i>trans</i>	Yes	Yes
NAWDEV_02	<i>trans</i>	Yes	Yes
NECZIE	<i>cis</i>	No	Yes
NECZUQ	<i>cis</i>	No	Yes
PIQPOU	<i>trans</i>	Yes	Yes
PIQPOU_02	<i>trans</i>	Yes	Yes
ROFHUP	<i>cis</i>	Yes	Yes
SEYPIW	<i>trans</i>	Yes	Yes
SEYPOC	<i>trans</i>	Yes	Yes
SUNPOG	<i>trans</i>	No	No
VAPNAB	<i>trans</i>	No	Yes
VEJWOW	<i>trans</i>	Yes	Yes
VUJCUY	<i>trans</i>	Yes	Yes
VUJCUY10	<i>trans</i>	Yes	Yes
WIVFOW	<i>cis</i>	No	Yes
WIVFOW_02	<i>cis</i>	No	Yes
XOTJAR	<i>cis</i>	Yes	Yes
XOTJEV	<i>trans</i>	No	Yes
YAPWIW	<i>cis</i>	No	Yes
YAPWOC	<i>cis</i>	No	Yes
ZIDWAC	<i>cis</i>	Yes	No

Table 2.2: Table of occurrences of the fragment: 3-chlorobut-2-ene-thiolate. The first column indicates the fragment reference in the CSD. The second column indicates whether the fragment is in *cis* or *trans* conformation. The third column indicates whether the atoms on the periphery of the fragment are constrained in some manner, such as in a five atom ring structure. The final column indicates if the sulfur atom is constrained within the original molecule or bonded to a hydrogen.

tions of the fragment 3-chlorobut-2-ene-thiolate based upon the assumptions made above. Since there are broadly three different possible conformations and there are two different options for each conformation, then it might be expected that there would be nine different combinations of conformations possible within these data. The fragments were analyzed by viewing fragments in Mercury [40] and assigning *cis/trans* and yes/no to either of the two other projected conformational changes predicted.

There are a number of problems with this approach. Principally, even before any structural analysis has taken place the user has a preconceived notion of what they would expect from the analysis. There is every possibility that assumptions made at this point may result in a researcher seeing what they want to see rather than what is actually there. Also, there is no geometric information gathered from this approach. While this may not be a serious problem with a simple example such as this, as the complexity of the fragment increases with the resulting increase in degrees of freedom within a large fragment these assumptions become more difficult to define. This is where a robust and universal definition of the geometry of the fragment becomes essential.

2.2.1.1 Total Geometries

The fragment 3-chlorobut-2-ene-thiolate was initially defined by total geometries and the results of these analysis are displayed in the Figure 2.2. In this figure the clusters are named A-G from left to right and the colours are carried from the dendrogram to the MMDS plot in the right of the figure.

In Figure 2.2, the clusters A-C are in the *trans* conformation and the fragments D-F are in the *cis* conformation. This is the biggest single conformational difference in these data. This is shown by the early separation of these fragments where the cut level joining these clusters has a low level of similarity. Within the fragments that are in *trans* conformation (A-C), the fragments that are in cluster C (green) are part of a five member ring system where the sulfur is part of this system. The integration into a five membered ring system results in the backbone of the fragment being constrained by the

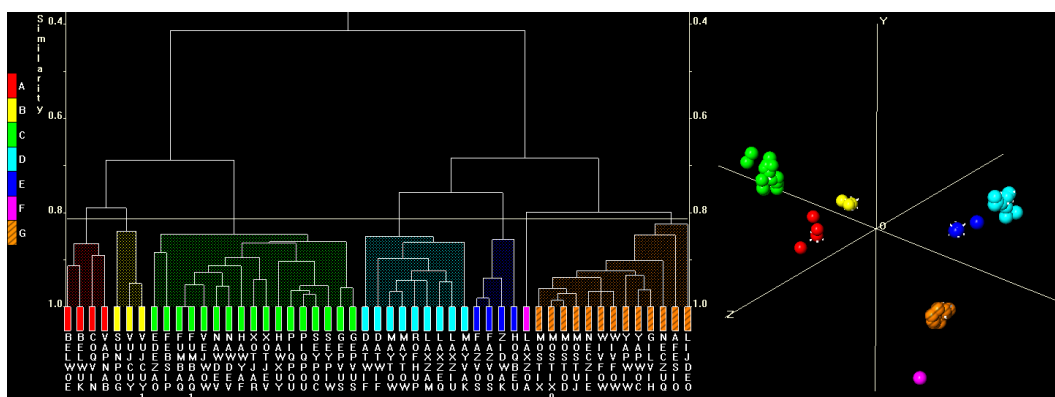


Figure 2.2: Figures indicating the differences in conformation in the fragment 3-chlorobut-2-ene-thiolate. The clusters are named A-G from left to right and the colours in the dendrogram are carried to the MMDS plot (right). The fragment was defined using total geometries.

chemical context that the fragment was derived from. The remaining *trans* fragments are either in a six member ring or in the backbone of the molecule. There is very little difference in conformation between clusters A and B. By examining the fragments in the fragment viewer, it appears that there is a minor twist around the double bond at the center of the fragment. The relative difference in structure between the two groups of fragments is small which is reflected in the close proximity of the fragments in the MMDS plot. The fragments that are in *cis* conformation have a similar distribution of conformations: The fragments that are in clusters D and E have the carbon backbone of the fragment in a five member ring system with the resulting restriction in bond angle. The difference in structure between clusters D and E is that the sulfur atom in cluster D is bound to a carbon atom in the molecule where the fragment was derived from. The fragments in cluster E have the sulfur atom bound to either another sulfur or a nitrogen atom. The chemical context of these fragments means that the bond angle around the sulfur atom has been altered in response to the environment that the fragment was derived from. The single fragment in cluster F is in an 8 member ring system while the fragments of cluster G have the carbon backbone of the fragment constrained in a 6 member ring system.

By viewing the fragments in Figure 2.3 it becomes clear what the different

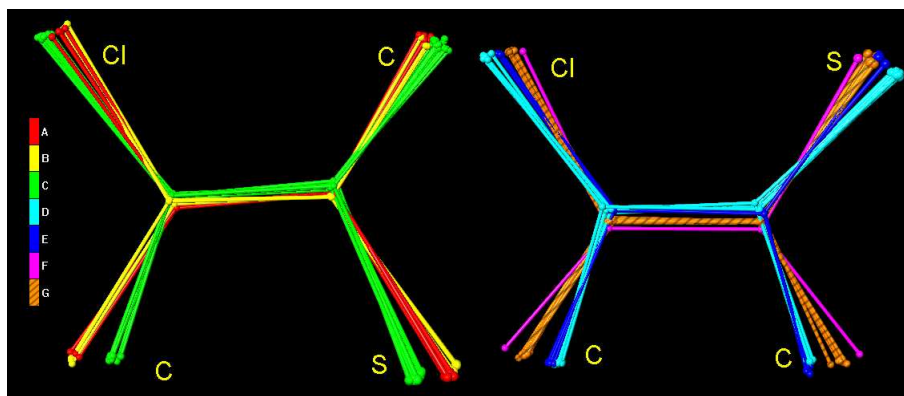


Figure 2.3: The overlay of the fragments 3-chlorobut-2-ene-thiolate. The fragments on the left are in *trans* conformation while the fragments on the right are in *cis* conformation. The fragments have been projected in such a way that the difference in bonded angles can easily be viewed for both conformations.

conformations of the fragments are. This combined with the dendrogram and MMDS plot in Figure 2.2 which displays the relationship in geometry between the fragments, makes it possible to very quickly determine what conformation the fragments are in, as well as getting an overview of the demography of the population. For instance it is possible to infer that approximately 50% of the population is in *trans* conformation and in this example a fragment in *trans* conformation is most likely to be found within a five membered ring system.

2.2.1.2 Clustering with angles only

The same analysis was carried out where the geometry of the fragments were defined by all angles only. These angles were all of the bonded and non-bonded angles. The results showed the fragments have been grouped into clusters that are similar to those formed when the fragments were defined with total geometries. That is, the fragments that are populating the groups when the geometry was defined with angles only are the same as those populating the groups when the fragments were defined by total geometries. The formation of similar clusters indicates that the distance matrix calculated

using Equation 1.1 was similar to the distance matrix that was calculated with total geometries. There is a single fragment that is in a different cluster. This fragment is SUNPOG which is found in the red cluster when the fragment was defined with angles only and in the yellow cluster when the fragment was defined by total geometries. When the fragments from both clusters are examined there is very little difference between the conformations of the fragment that populate these clusters. It could therefore be regarded as a minor rearrangement in classification of the fragments within these clusters. Within the remaining clusters there are only minor difference in the relationship between fragments. For example, the green cluster when the fragments are defined by both total geometries and angles only, contain the same fragments in both cases. By examining the fragments in Figure 2.3 the difference in conformation of the green fragments are minor. There was one notable difference however; all of the fragments have a greater similarity and this change was manifested by the dendrogram being ‘shorter’ than that of the dendrogram calculated using the distance matrix calculated using total geometries. The shorter dendrogram indicates that the fragments appear to have similarity that the same fragments defined by total geometries. This indicates that overall the fragments appear to have greater similarity than when the fragments were defined by total geometries. Comparing the MMDs plots in Figures 2.2 and 2.4. Both of the plots show that the fragments are grouped into isolated groups. This shows that either of these definitions can adequately describe the conformations of the fragments in this case.

2.2.1.3 Clustering with distances only

When cluster analysis was applied to distances only, the reverse was seen. In the above example when the fragments were defined with angles only, these fragments appeared to be have greater similarity. In the case of distances only the fragments appear to be less similar. The dendrogram was ‘taller’ indicating that the fragments were less similar to each other than when total geometries were clustered. Examining clusters A, B and C when the fragment was defined by all distances only, contain the same fragments that make up

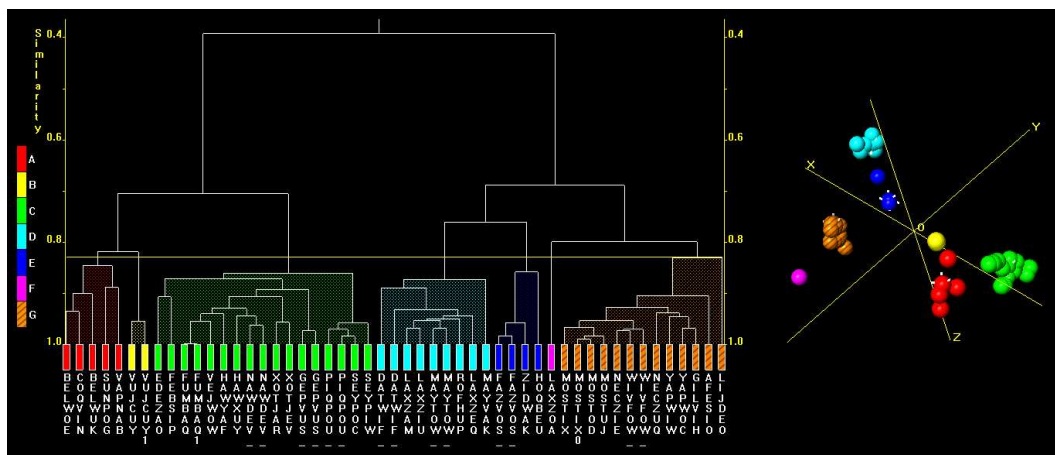


Figure 2.4: Dendrogram and MMDS plot clustered using all angles only for the fragment 3-chlorobut-2-ene-thiolate.

clusters A and B when the fragments were defined by total geometries. As stated previously, these fragments are quite similar in conformation. This can be seen in Figure 2.3. There has been a merging of the fragment LAXZOA into the orange striped cluster and the fragment AFESIO has been formed into a separate cluster. When these fragments are examined in the fragment viewer there is very little difference between the fragments in these clusters. This indicates that the distance matrix calculated when the fragment was defined with distances only is similar to the distance matrix calculated when the fragment was defined by total geometries. When the MMDS plots are compared between Figures 2.2 and 2.5, it shows that the groups are still isolated from one another indicating that the fragments in these groups share similar conformations. It should be noted that the clusters in Figure 2.5 are more diffuse than when the geometry of the fragment was defined by total geometries. This could be symptomatic of the fragments appearing to be more dissimilar when the fragments were defined using distances only.

2.2.1.4 Bonded distances and angles only

Cluster analysis was performed on the same fragment with just the bonded distance and bonded angles defined. The dendrogram generated from the distance matrix was radically different to that of the previous three analy-

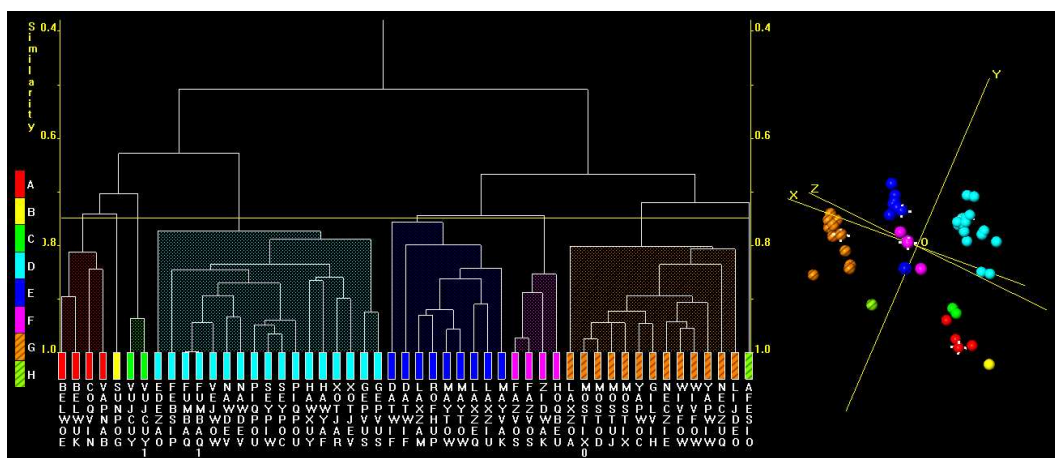


Figure 2.5: Dendrogram and MMDS plot clustered using all distances only for the fragment 3-chlorobut-2-ene-thiolate.

ses. The dendrogram can be seen in Figure 2.6. There was no distinction between *cis* and *trans* conformation and there is considerable reshuffling of the fragments both within clusters and within the sample data itself. An interesting observation is that using this definition there are actually clusters of fragments that have an underlying structural basis. The fragments appear to have been grouped into clusters according to the constraints placed around the double bond, that is the fragments that are found in a five membered ring system are found within the same cluster. While this is useful there is still only limited conformational information retrieved by this analysis. The major conformational changes have not been detected by this geometric definition. The MMDS plot in Figure 2.6 shows that the clusters are quite diffuse. This suggests that using this definition the structural changes detected using this definition has not been accurately described. Ideally a single cluster of fragments in the MMDS plot should be grouped together in close proximity. It would be expected that for a fragment with a discrete conformations accurately described should form isolated continuous clusters in the MMDS plot.

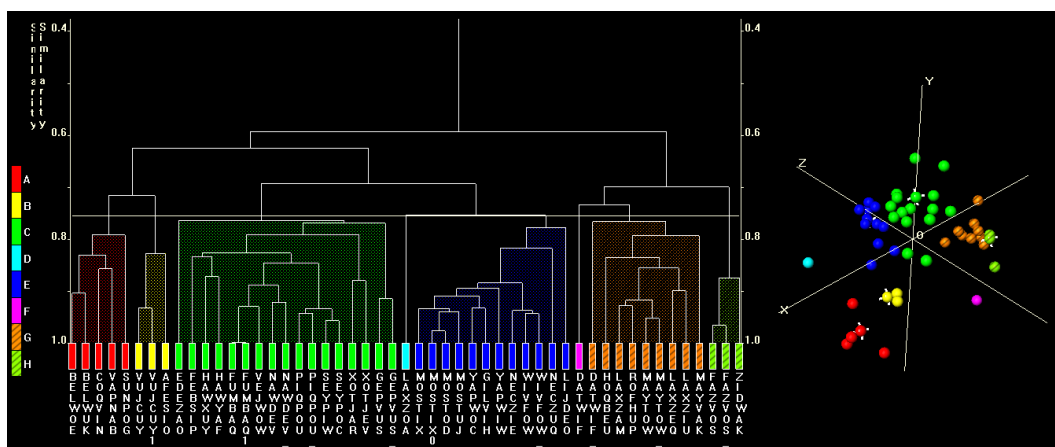


Figure 2.6: Dendrogram and MMDS plot clustered using bonded angles and bonded distances only for the fragment 3-chlorobut-2-ene-thiolate.

2.2.1.5 Bonded distances, angles and backbone torsion angle

The dendrogram in Figure 2.7 was generated when the fragment was defined by bonded angles and distances and a single torsion angle defining the carbon backbone. The addition of this torsion angle was intended to differentiate between the different conformations, particularly the *cis/trans* conformational change. When the results were examined it appeared that the fragments have formed clusters that contain fragments with the same conformation. This is illustrated in Figure 2.8. This figure shows the fragments that have been separated into *cis* and *trans* where the fragments have been coloured according to the colour of the fragments in Figure 2.7. By examining the fragments it should be noted that where the backbone of the fragment had been constrained these fragments have formed clusters. By examining the MMDS plot it is possible to see that these clusters are isolated but are more diffuse than when the fragment was defined by total geometries. Ideally the fragments should be in isolated and continuous clusters.

2.2.2 3-aminobutan-2-ol

The next fragment that was studied was 3-aminobutan-2-ol and the search criteria is described in Table 2.1. There are two different major conforma-

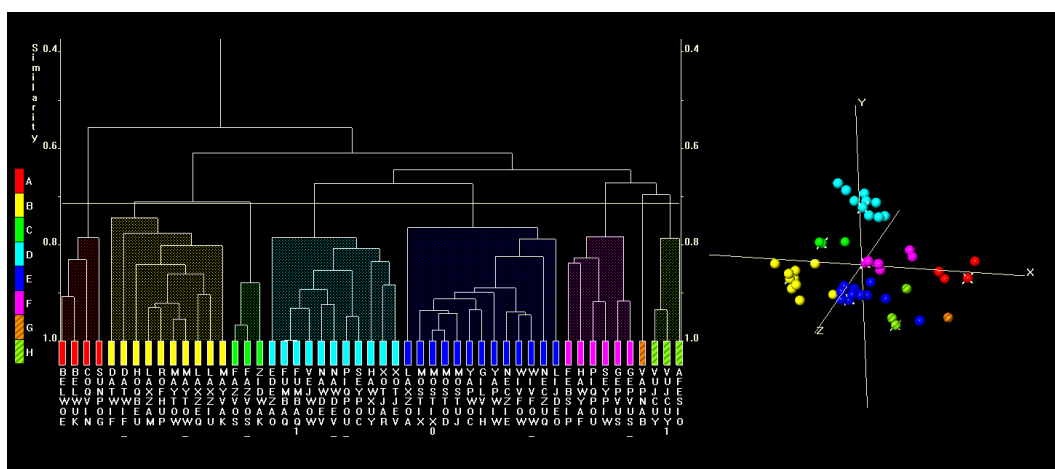


Figure 2.7: Dendrogram and MMDS plot clustered using bonded angles, bonded distances and carbon backbone torsion angle for the fragment 3-chlorobut-2-ene-thiolate.

tional changes in this dataset. These are illustrated in Figure 2.10 and consist of a rotational component and a restriction in the bonded angles as a result of the chemical context that the fragment was derived from. In Figure 2.10, the hydrogen atoms are not illustrated but are specified in the search and as a result the fragments have two chiral centres. Since the central bond in the fragment is a single (σ) bond there will be the typical steric hindrance associated with two sp^3 carbon systems interacting with each other. That is that the *staggered* conformation is more energetically favourable than the *eclipsed* conformation and the *anti* conformation is more energetically favourable than the *gauche* conformation (Figure 2.9). While this may be the case from a molecular perspective, from a fragment perspective, the conformations that each fragment can undertake will vary to a greater degree owing to the context that the fragment finds itself in. That is, the fragment could be in an energetically favourable conformation if that is energetically favourable for the entire molecule.

2.2.2.1 Total Geometries

Initially the fragment was defined by total geometries. This definition comprises of all angles and distances including both the bonded and non bonded

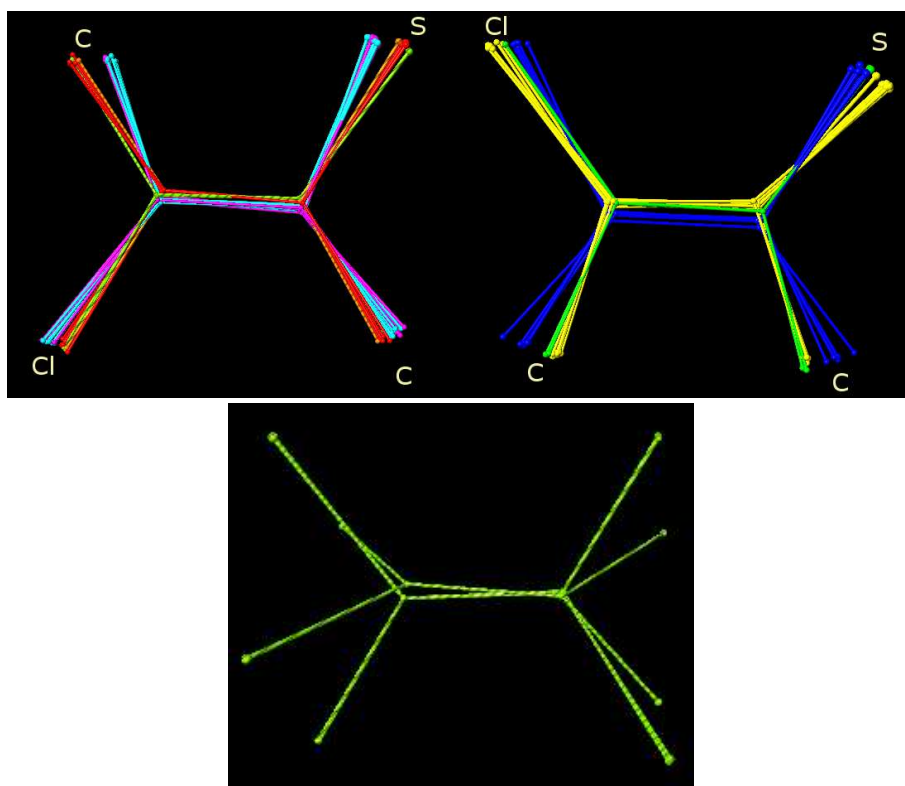


Figure 2.8: Fragment overlay of 3-chlorobut-2-ene-thiolate where the fragment was defined by bonded angles, bonded distances and backbone torsion. The fragments have been separated into *cis* on the right and *trans* on the right. Below is the cluster H where the fragments are not in a typical conformation around a double bond.

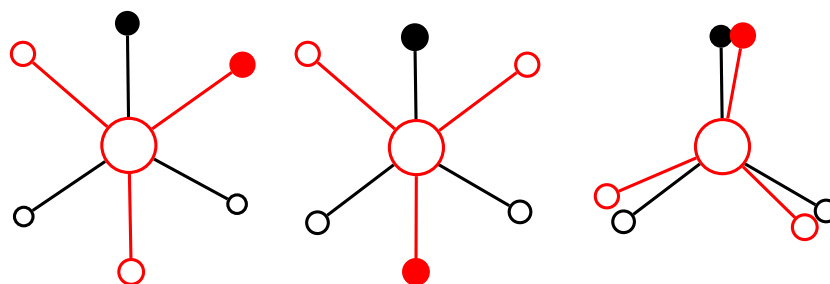


Figure 2.9: An example of Newman projections. The left hand projection is in the *anti* conformation, the center projection is in the *gauche* conformation. Both of these conformations are staggered while the right hand projection is in eclipsed position.

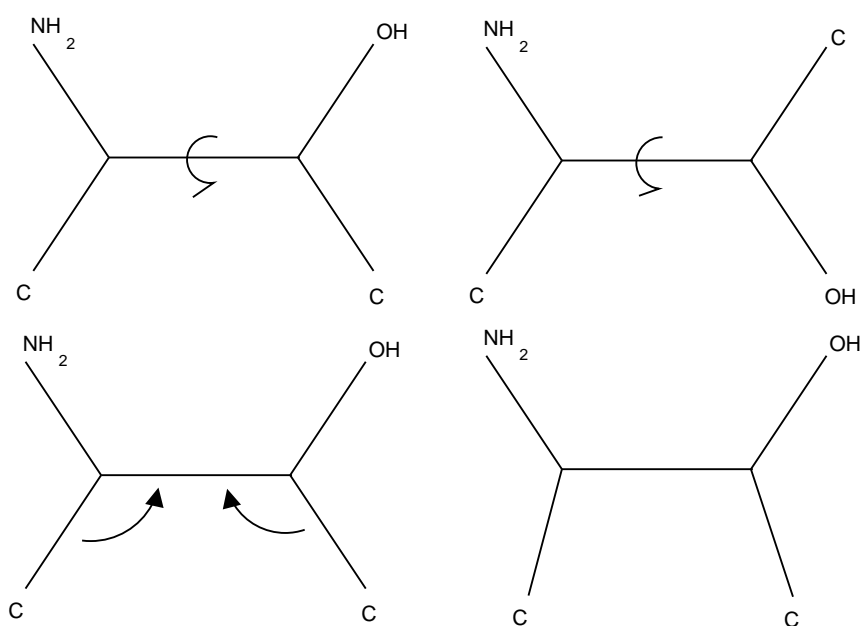


Figure 2.10: Predicted conformations of the fragment 3-aminobutan-2-ol. Top is an indication of the rotational change around the central bond. Bottom is an illustration of the constraint that can be placed on the backbone of the fragment by the original molecule from where the fragment was derived. Also there are two chiral centres which could be found in S^* and R^* conformation.

variables. These variables are all scalar values. This results in the analysis being unable to distinguish between fragments of different absolute chirality. [10, 15] These fragments will fall into the same cluster. As discussed in Collins *et al* [15] the absolute configuration may be disregarded in some cases. It is also the case that the absolute configuration of the fragments may not have been determined in the CSD [24]. Using the CSD, it is possible to select a specific configuration but this can reduce the number of hits in the database[15]. In this case the absolute configuration has been ignored. An example of this are the fragments in cluster A (red) in Figure 2.11. Within this cluster the relative chirality of the fragments appear to be S*-S* and R*-R* respectively. It should be noted that it is possible to detect the relative differences between chiral centres. When exact enantiomers, where 2 fragments are identical mirror images of each other, are examined the distances between atoms are equal. Thus, scalar variables will be unable to differentiate between exact enantiomers. This means that the rotational conformation about the central bond is the most important change in these data. The diagram in Figure 2.11 gives an illustration of the rotational conformational of the fragments that are found in each of the clusters.

The fragments have been clearly clustered into groups that have different conformations. The rotational conformation of these data is illustrated by the Newman projections in Figure 2.11. The fragments in clusters B and C have the same conformation but the fragments that have been grouped into cluster B are constrained within a five atom ring structure. This indicates that the conformations of these fragments are accurately defined by this definition. That is, the clusters of fragments generated using this definition are discrete in conformations. This is best illustrated using the MMDS plot in Figure 2.12 where all of the fragments are in isolated clusters with the sole exception being the fragment OJUYUN (Orange striped) which has a very similar conformation to the fragments in the green cluster (C) but is distinct since the all of the bond lengths appear to be shorter.

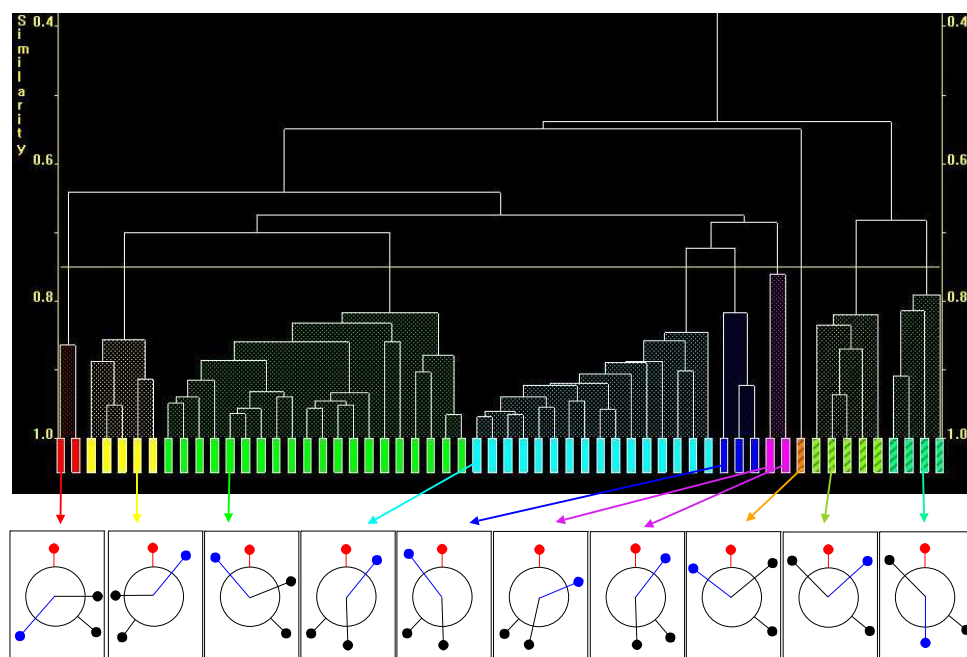


Figure 2.11: Dendrogram and Newman projections of the fragment 3-aminobutan-2-ol. The geometry of the fragments in this dendrogram have been defined by total geometries. The Newman projections represent the conformation of the fragment within that cluster.

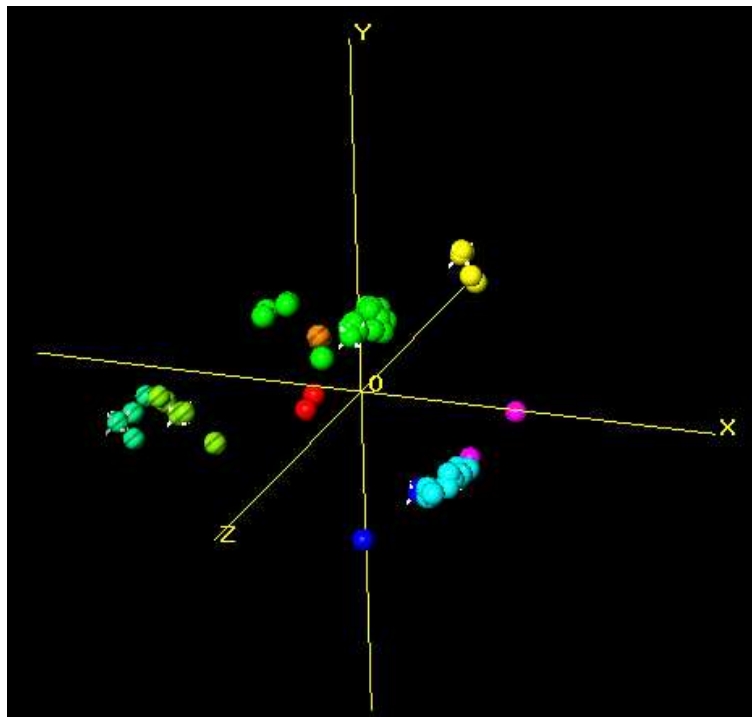


Figure 2.12: MMDS plot clustered using total geometries for the fragment 3-aminobutan-ol.

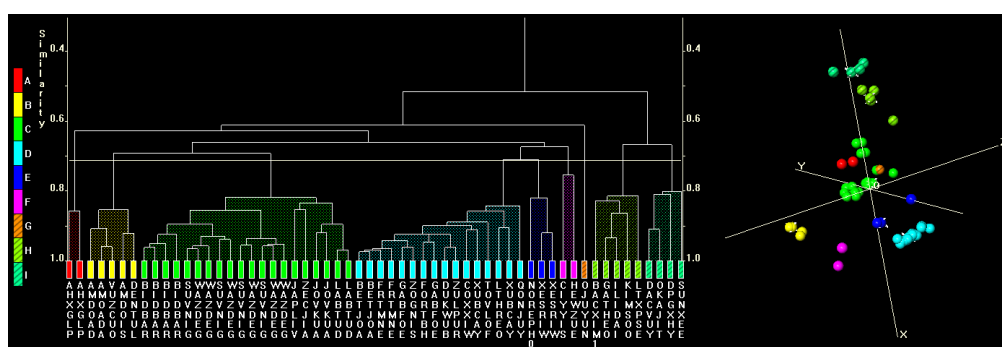


Figure 2.13: Dendrogram and MMDS plot clustered using all angles only for the fragment 3aminobutan-2-ol. The cut level was set at 0.814

Cluster name (Total Geometries)	Cluster equivalents (Angles)
A	A
B	B
C	C*
D	D*
E	G
F	F
G	E
H	H
I	I

Table 2.3: Cluster equivalents; clusters are compared between those fragments defined by angles only and those determined by total geometries. *Denotes minor rearrangement within the cluster

2.2.2.2 Angles only

When the geometry of the fragments were defined by angles only, the dendrogram and MMDS in Figure 2.13 was generated by *d*SNAP. This output was examined and the population of the clusters was compared with the clusters formed when the fragments were defined by total geometries. The comparisons are tabulated in Table 2.3. This table indicates that when the fragment was defined by angles only, the fragments were grouped into the same clusters as when the fragment was defined by total geometries. The fragments in these clusters differ only by the manner in which the fragments are related. This shows that the definition of all angles only has described the conformation extremely well. Even the fragment OJUYUN (Orange striped) which is an unusual conformation where the bond lengths are shorter than the fragments in a similar rotational conformation (Cluster C) is in an isolated cluster. The differences in the tie bars between the dendrograms where the fragments were defined by total geometries and all angles only can be explained by minor differences in the distance matrix. It should be noted that the classification of the fragments was not effected by this minor difference in distance matrix.

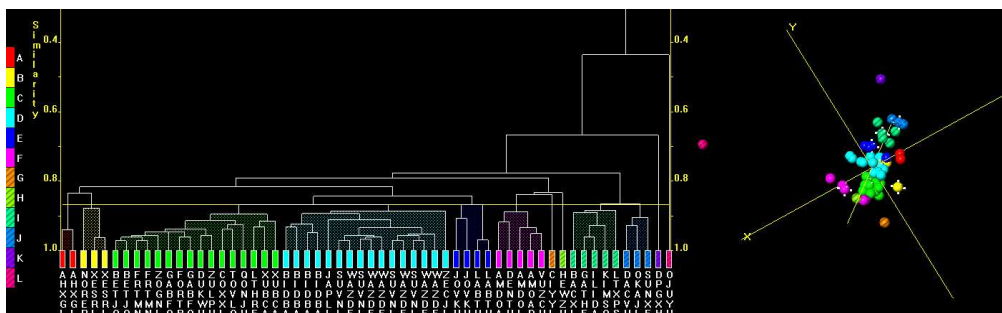


Figure 2.14: Dendrogram and MMDs plot clustered using all distances only for the fragment 3-aminobutan-2-ol.

2.2.2.3 Distances only

In this example, the geometry of the fragment was defined by all distances only and the output is shown in Figure 2.14. There is a strong resemblance between the dendrogram in this figure and the one in Figure 2.11. This similarity indicates that when cluster analysis was carried out with these data, the definition used in this example has accurately described the shape of the fragment. Table 2.4 gives a comparison between the clusters that have formed when the fragments were defined by total geometries and by atomic distances only. As can be seen in Table 2.4 some of the clusters have been split. An example of this is the green cluster (cluster C) where the fragment is defined using total geometries has been split into two clusters (cyan and blue, clusters D and E) when the fragment was defined using all atomic distances. The difference between these clusters when defined by all atomic distances only, is a difference between the rotation around the central carbon bond. Other than this, the conformation of the fragments are very similar. The fragments in both the cyan and blue clusters have the same relative chirality and very similar bond lengths and bond angles. The fragment OJUYUN (pink striped in Figure 2.14) differs from the other fragments as the bonded distances are shorter. Using all atomic distances only to describe the conformation of the fragments results in this fragment appearing to very different to the rest of the dataset. There is a similar but much less dramatic reason for the the split of cluster I (light green striped)

Cluster name (Total Geometries)	Cluster equivalents (Distances)
A	A
B	F*
C	D & E†*
D	C*
E	B
F	G & H†
G	L
H	I
I	J & K†

Table 2.4: Cluster Equivalents: Total geometries against distance * denotes that there has been rearrangement within the cluster. † denotes that the cluster has been split into 2 clusters.

when the fragments was defined by total geometries has split into two clusters J and K (blue striped and purple striped). The reason for this split is a small difference in bond length in the fragment DPGXHY which is found in cluster K when the fragment was defined by all atomic distances only. Inspire of these differences in the clusters when using all atomic distances compared to total geometries, the definition has organised the fragments into groups that have rational structural reasons underpinning their formation. It is noted that some of the differences are the result of differences in bond lengths of some of the fragments which is exaggerated when defining the geometry of the fragments with distances only.

2.2.2.4 Bonded distances and angles

When the geometry of the fragment was defined using only the bonded variables, the resulting dendrogram is shown in Figure 2.15. The output has very little in common in appearance with the dendrogram in Figure 2.11 where the fragment was defined by total geometries. By examining the fragments in the red cluster in Figure 2.15 it is apparent the this cluster is made up of many fragments of different conformations when viewed in the fragment viewer. When these fragments are located when the fragment was defined by total geometries is it clear that the red cluster is made up of a mixture of

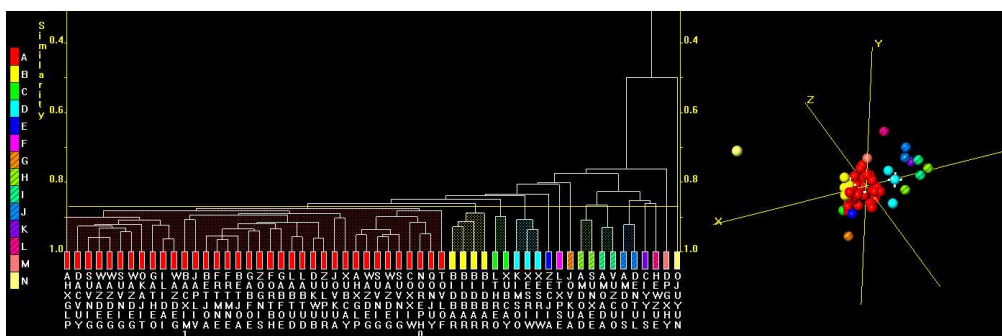


Figure 2.15: Dendrogram and MMDS plot clustered using bonded angles and distances only for the fragment 3-aminobutan-2-ol. This is a good example of the output from *d*SNAP where the clusters have been poorly defined.

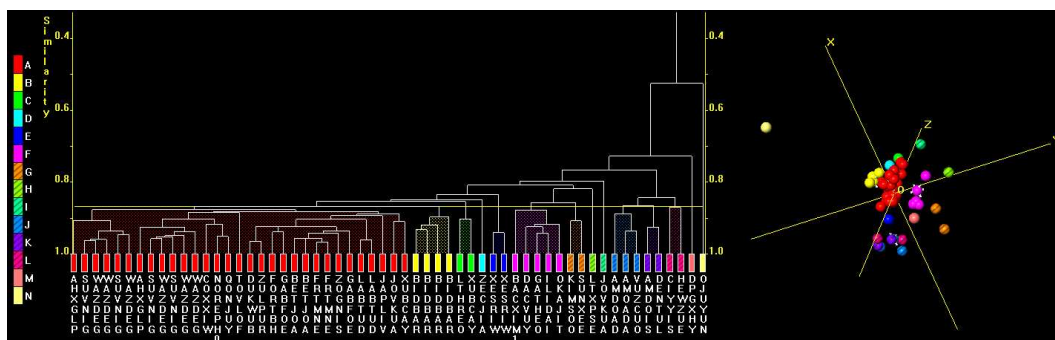


Figure 2.16: Dendrogram and MMDS plot clustered using bonded angles, bonded distances and backbone torsion for the fragment 3-aminobutan-2-ol.

clusters A, C, D, E, H and I. Since the fragments that make up these clusters are made up of fragments in discrete conformations when described by total geometries, it is clear that using only the bonded distances and bonded angles fails to accurately describe the geometry of the fragments.

2.2.2.5 Bonded distances, angles and backbone torsion

In this section a single torsion angle between atoms C1, C2, C3 and C4 as seen in Table 2.1, is added to the analysis where the fragment was defined by bonded distances and angles only. This is an attempt to resolve the deficiencies that were found in the above section. It was hoped that the addition of the torsional information will aid in the distinction between the

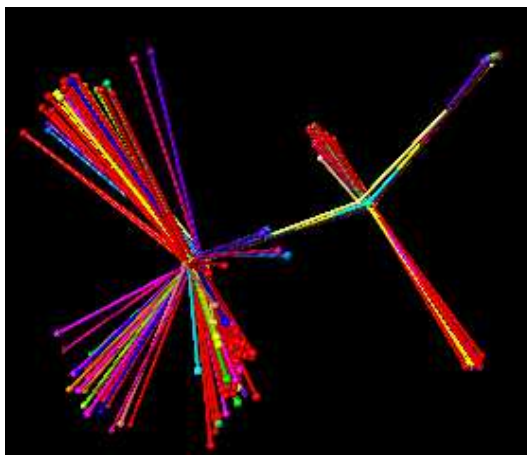


Figure 2.17: An overlay of the fragment 3-aminibutan-2-ol where the fragment was defined by bonded variables and backbone torsion. The fragments in this figure have been coloured to correspond with the colours in the dendrogram. As can be seen, the fragments have not formed clusters that have a consistent conformation.

different rotational conformations in these data. Unfortunately, when the output from this definition is compared with the output from total geometries there does not appear to be any similarity between the outputs. In Figure 2.16 the red cluster is made up of a collection of fragments that were in discrete clusters when the fragment was defined by total geometries. This can be illustrated when the fragments are examined in the fragment viewer. With reference to Figure 2.17 it is apparent that there is no conformational reason for the formation clusters.

2.2.3 Pentan-2-one

The final example is pentan-2-one and is illustrated in Figure 2.18 and essentially there are two independent torsional rotations that are indicated by the blue arrows in Figure 2.18. The carbon backbone of this fragment was restricted to acyclic bonds when the search in the CSD was carried out in order to keep the bonded angles as consistent as possible throughout the data set. This restriction aims to ensure that the rotational conformational changes will dominate these data and reduce the number of hits in the database to a

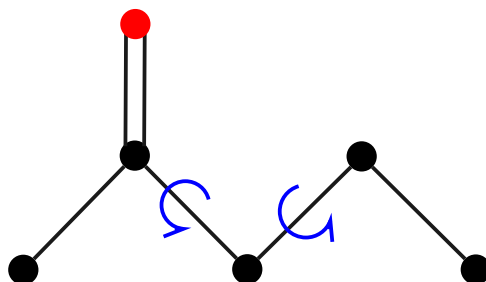


Figure 2.18: Diagram of the fragment pentan-2-one with the free torsions indicated by the blue arrows.

more manageable level. The search criteria and number of fragments can be seen in Table 2.1.

2.2.3.1 Total Geometries

As can be seen in Figure 2.19, each of the clusters that have formed in the analysis can be attributed to a specific conformation of the fragments. Figure 2.21 illustrates the different conformations that the fragments have taken. Each of the fragments has formed a distinct conformation that correspond to the Newman projections in Figure 2.19. When the fragments are overlaid using the fragment viewer it is clear that the fragments have broadly classified into discrete conformations. This can be seen in Figure 2.21. In this figure the fragments have been aligned in such a way that the three atoms in the background of the figure have been superimposed. These atoms are the atoms of the ketone group and as a result the rotational nature of the backbone are exaggerated. As illustrated by the Newman projections in Figure 2.19 the differences between the red and yellow clusters in this figure and the cyan and blue clusters is a difference in the torsion angle closest to the ketone group. The green cluster is distinct from the other clusters but in terms of the torsion angle but is closest in conformation to the yellow cluster. This similarity is illustrated by the dendrogram in Figure 2.19 and the MMDS plot in Figure 2.20. The second torsion angle in this fragment differentiates between the red and yellow clusters and the cyan and blue clusters. Again, by examining the fragments in the fragment viewer in Figure 2.21

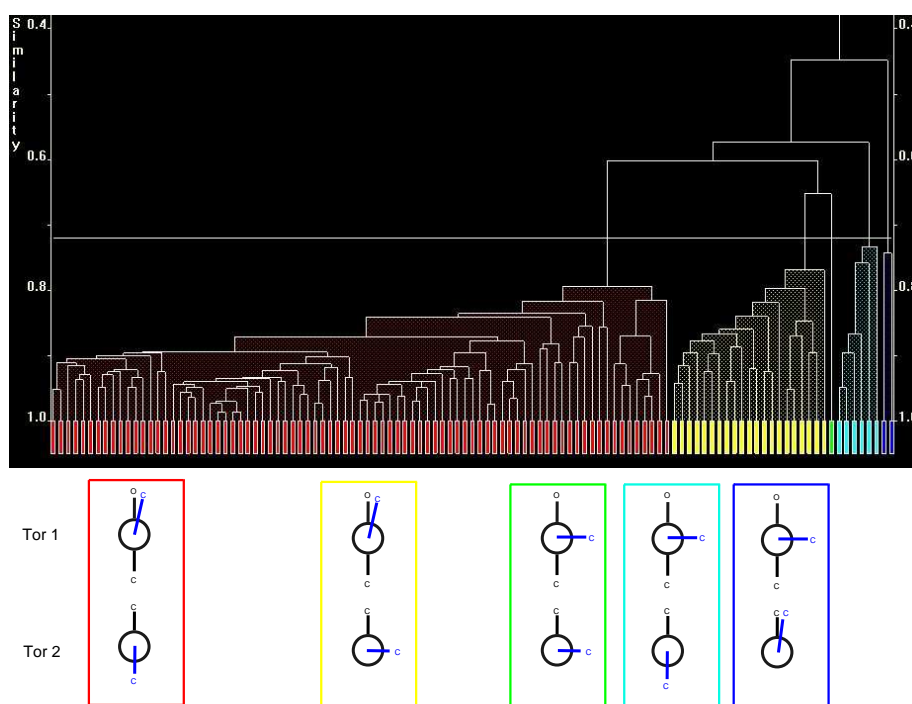


Figure 2.19: Dendrogram clustered using total geometries for the fragment pentan-2-one. The Newman projections on the lower part of the diagram reflect the torsional rotation of the two torsion angles.

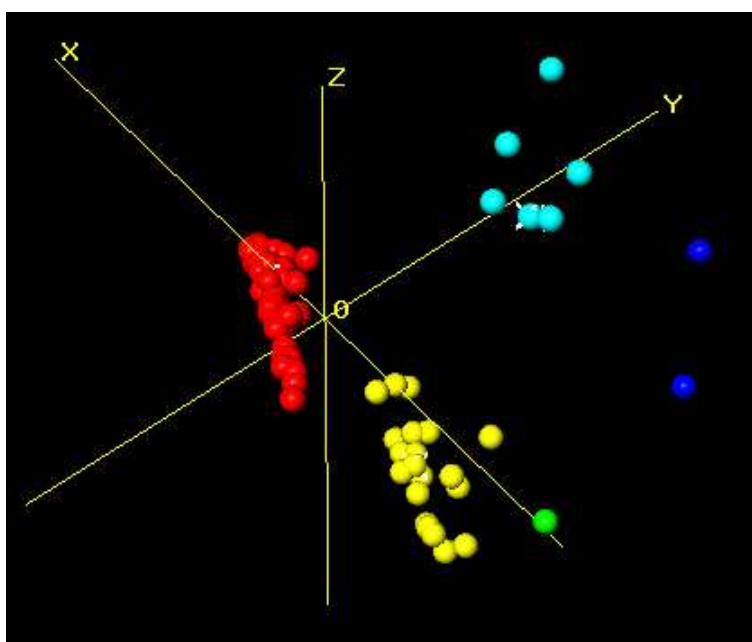


Figure 2.20: MMDS plot clustered using total geometries for the fragment pentan-2-one. The colours of the spheres representing the fragments have been taken from the dendrogram in Figure 2.19.

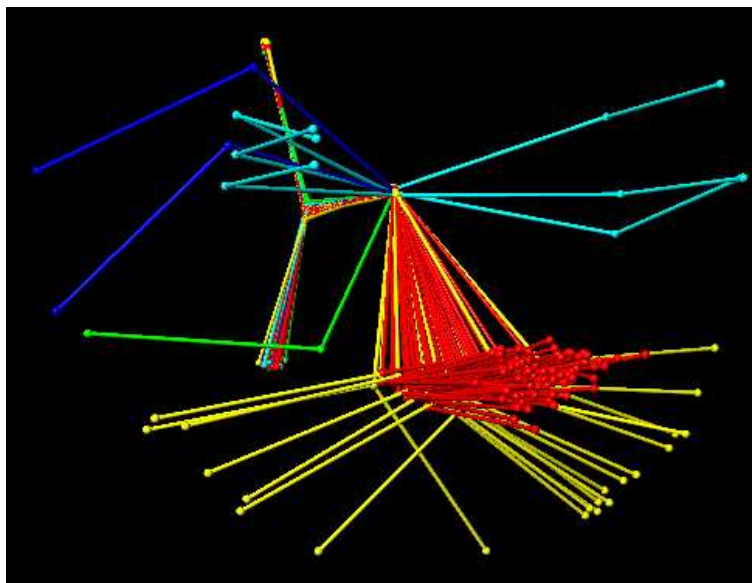


Figure 2.21: Overlay of the fragment pentan-2-one where the geometry was defined by total geometries. The overlay provided by the fragment viewer shows that the fragments are grouped into clusters with similar conformations.

and the Newman projections in Figure 2.19 it is clear that the difference in the value of the second torsion angle is causing these clusters to form. This indicates that when the fragments are defined using total geometries cluster analysis has successfully grouped the fragments according to the conformation of the fragments. It should be noted that because of the scalar nature of total geometries it is impossible to differentiate between mirror images of fragments.

2.2.3.2 Angles only

The dendrogram in Figure 2.22 was generated when the fragment was defined by all angles only. The dendrogram was compared with the dendrogram generated when the fragment was defined by total geometries (Figure 2.22). There are a few differences in relationship between the each of the clusters. For example, when the fragment was defined by total geometries the yellow cluster (B) has become the cyan cluster (D) when the fragment was defined

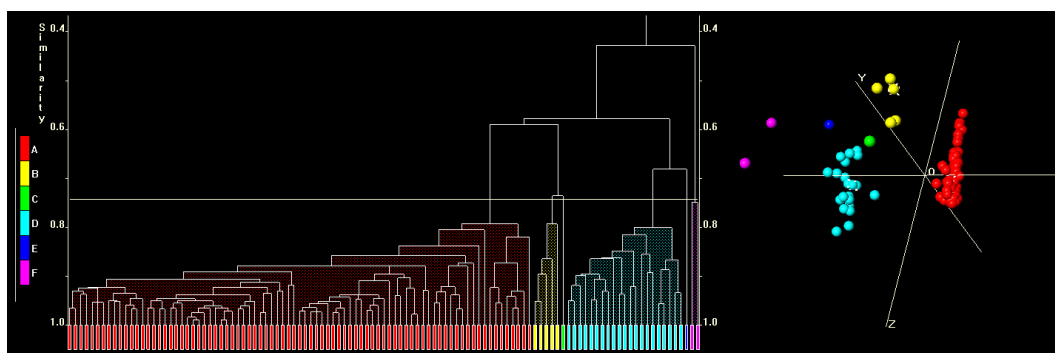


Figure 2.22: Dendrogram and MMDS plot clustered using all angles only for the fragment pentan-2-one.

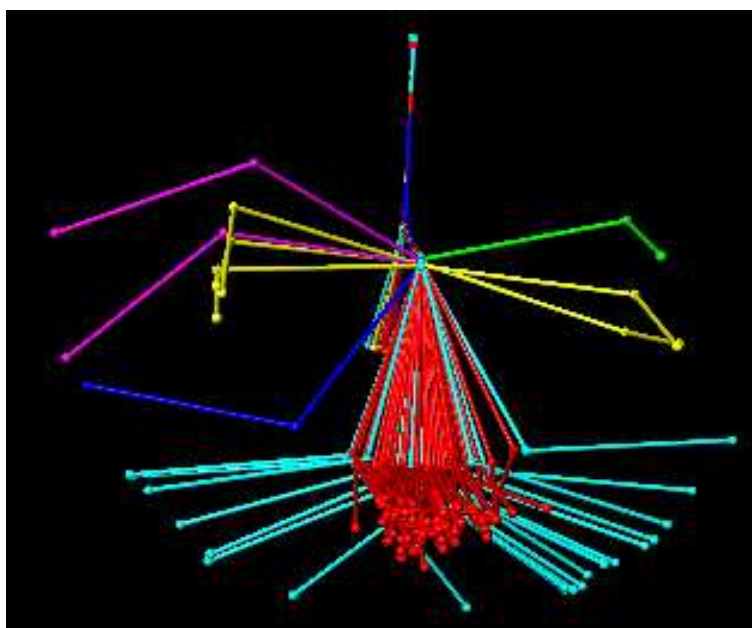


Figure 2.23: Overlay of the fragment viewer 3-aminobutan-2-ol when the fragment was defined by angles only.

by total geometries. These differences are tabulated in Table A.1 on page 138. This table shows which cluster each fragment falls into for each of the geometric definitions. This table shows that when the fragments had their geometry defined by angles they form almost identical clusters when compared to the clusters formed when the fragments were defined by total geometries. The only difference is the fragment FIVGOG which is in a cluster of its own. Examining the green fragment in the dendrogram in Figure 2.22 and in Figure 2.23 it is clear that this fragment is closely related in conformation to the yellow clusters.

It should be noted that the biggest difference between these two definitions is the different relationship between clusters not by the members of those clusters. By examining the dendrogram in Figure 2.22 it is clear the cyan cluster less closely related to the red cluster which is indicated by the higher tie bar between these clusters. When geometry of the fragments were defined using total geometries (Figure 2.19), the yellow cluster is the same as the cyan clusters. As can be seen in Figure 2.19 the red and yellow clusters are more similar when the fragments are defined by total geometries than when the same fragments are defined by angles only. The reason for this difference in relationship is because the fragments have been grouped according to similarities in the second torsion bond which is furthest from the keto group in this fragment. When the fragments have been clustered using angles only, the fragments that have similar conformations for the second torsion angles are more closely related than when the fragments have been defined by total geometries. This difference in relationship does not mean that the fragments have been misclassified from the perspective of their conformations. All of the fragments, with the exception on FIVGOG, are found in isolated clusters of distinct conformations. This shows that the definition of angles only can, in this case, accurately classify the geometries of fragments into clusters.

2.2.3.3 Distances Only

The dendrogram in Figure 2.19 where the fragment was defined by total geometries is compared with the dendrogram in Figure 2.24 where the fragment

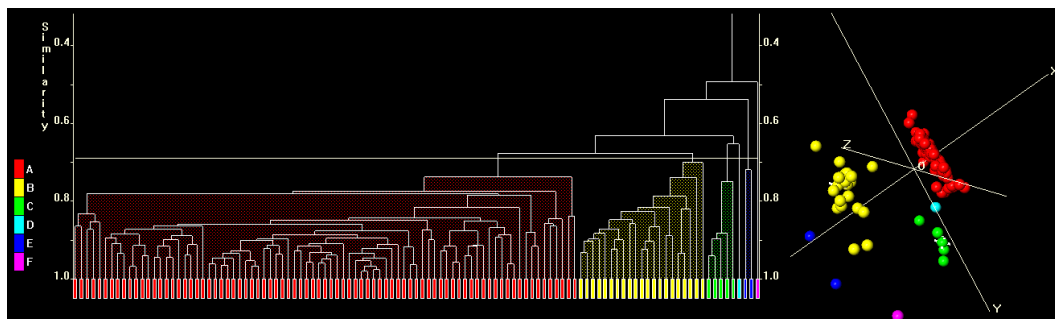


Figure 2.24: Dendrogram and MMDS plot clustered using all distances only of the fragment pentan-2-one.

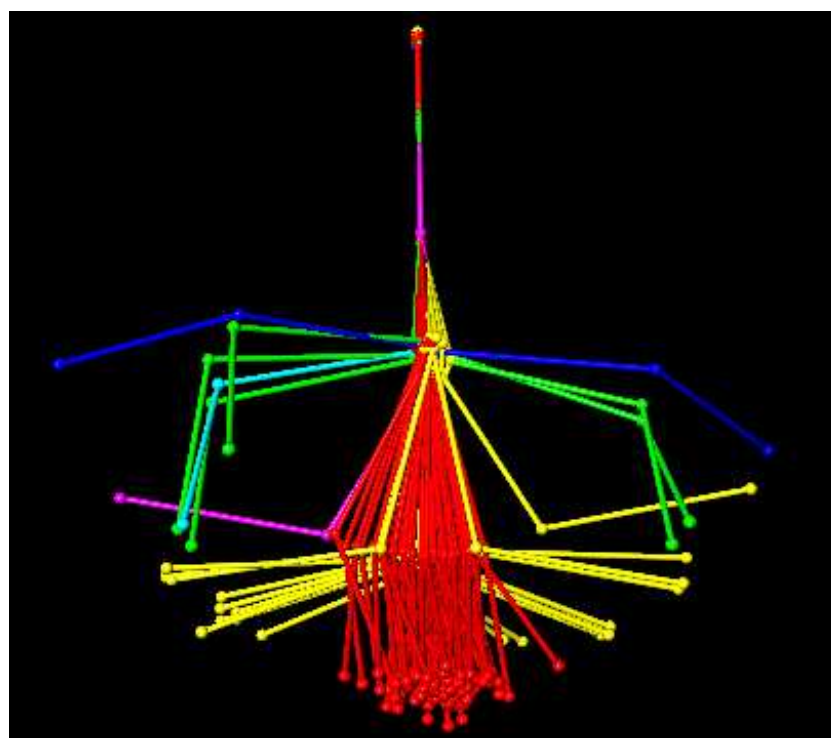


Figure 2.25: Fragment view of fragment pentan-2-one when the geometry of the fragment was defined by atomic distances only

was defined with atomic distances only. There appears to be good agreement between the output from *d*SNAP when the fragments were defined by total geometries (Figure 2.19) and atomic distances only. Most of the fragments have been grouped together with fragments of similar geometries. This has been tabulated in Table A.1 on page 138.

This table shows that the fragments have been grouped into clusters with good agreement with the clusters formed when the fragment was defined with total geometries. The whole of the red and yellow clusters have been preserved when the fragments was defined by distances only and total geometries. Also, the fragments that are in both of these clusters are made up of fragments of discrete conformation in both definitions. The green and cyan clusters when the fragments were defined by distances only, contain the fragments from the cyan clusters when the fragment was defined by total geometries. The apparent anomaly is only the result of the cut level of the dendrogram. If the cut level was set higher then this anomaly would disappear. Unfortunately, if the cut level was set higher then the rest of the clusters would not make sense. The remaining three fragments are the blue and green fragments when the fragments was defined by total geometries. The green fragment (MERWIQ) when the fragment was defined with total geometries is the purple fragment when the fragment was defined using distances only. This fragment is now regarded as very different when defined by all distance only. This is a result of some of the bond lengths being different than the rest of this dataset. Bonds C1-C2, C2-C3 are shorter and C2-O6 is longer. The C3-C4 and C4-C5 bond length very similar to the rest of the dataset. The blue fragments when the fragment is defined distances only is the same blue cluster when the fragment was defined with total geometries. Overall, defining the fragment with all distances only has successfully classified the fragments into groups that have a clear structural rational underpinning the clustering. The only exception is the fragment MERWIQ whose differences can be explained by unusual bond lengths in this particular fragment.

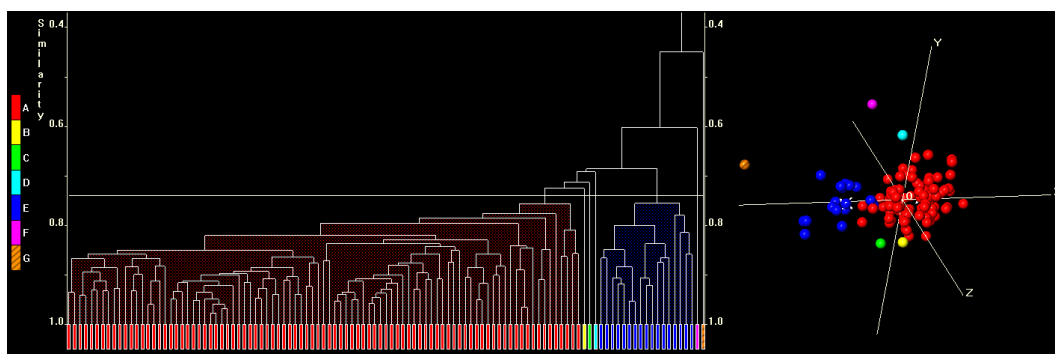


Figure 2.26: Dendrogram and MMDS plot clustered using bonded variables only for the fragment pentan-2-one.

2.2.3.4 Bonded Variables

Looking at the dendrogram in Figure 2.26, it is apparent that visually the dendrogram shows little resemblance to the dendrogram in Figure 2.19 where the fragments were defined using total geometries. By examining the red cluster in the fragment viewer (Figure 2.27) it is apparent that this cluster is made up of fragments of many different conformations. This indicates that the fragments have not been accurately grouped into fragments according to their conformation. Nevertheless there appears to be two distinct groups of fragments (red and blue) along with a number of isolated fragments. This difference is a result of subtle differences in bonded angles between the keto group and the carbon backbone that is not entirely obvious in the initial analysis where the fragments were defined using total geometries. The remaining fragments that are not in the red and blue clusters are different from these clusters as the result of different bond lengths. Overall, when the fragment was described using all bonded variables has failed to accurately differentiate between the different conformation of fragments present in these data. This would indicate that using all bonded variables as a description of the geometry of fragments is inappropriate for this fragment.

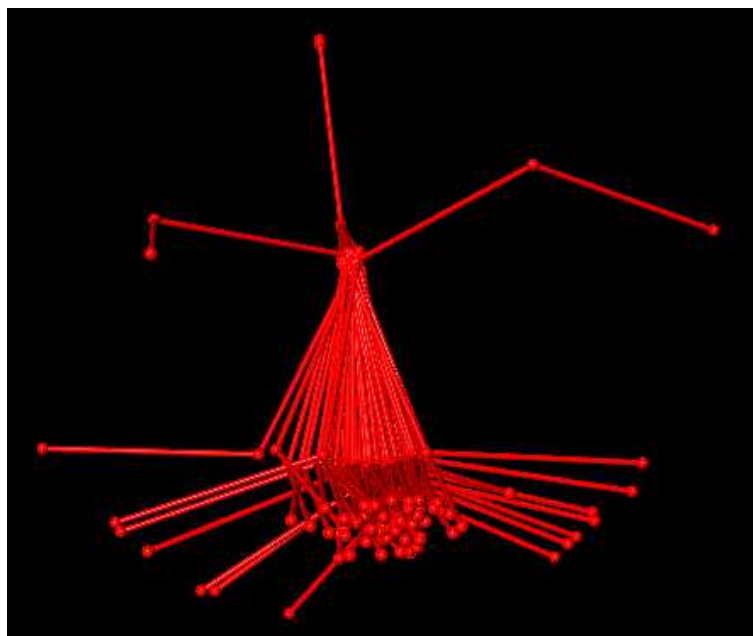


Figure 2.27: Fragment view of cluster A where the fragment was defined by all bonded variables only.

2.2.3.5 Bonded variables and torsion

An attempt to add two backbone torsion angles to the bonded variables failed. This was due to an inability to correct for the handedness of the torsion. That is, by convention a torsion is described as an angle between -180° to $+180^\circ$ but *dSNAP* can only accept positive scalar values as input. If the torsion is corrected such that the angle varies between 0° and 360° there is still a problem when comparing a torsion with an angle of 1° with one of 359° . Using the clustering algorithm in *dSNAP* the difference between these two variables will be 358° . This is incorrect from a chemical perspective as the difference should be a 2° difference in torsion angle.

2.3 Conclusions

There appears to be a significant difference between the different types of definition: definitions that include non-bonded information and those without. When the fragments were defined by bonded distances and angles only, there

is not enough structural information described by these variables to generate clusters of fragments into groups with similar geometry. This definition did not contain any non-bonded information and this appears to be the reason why there is no correlation between the clusters and the conformation of the fragments grouped into those clusters.

When torsion angles were added to the analysis it was hoped that this deficiency would be remedied and the fragments would be grouped into cluster that could be justified from a chemical perspective. Unfortunately this is not the case. In the example of 3-aminobutan-2-ol when the carbon backbone was defined with a torsion angle there was no distinction between the different rotational conformations of the fragments and the formation of clusters. This is illustrated when the dendrogram in Figure 2.16, where the fragment was defined with bonded variables and a backbone torsion angle, is compared to the dendrogram in Figure 2.11, where the fragment was defined with total geometries. When the contents of the clusters in Figure 2.16 are examined, there does not appear to be any relationship between the clusters and the geometry of the fragments within those clusters.

It would appear that it is necessary for the variables describing the geometry of the fragments to have non-bonded information. When the fragments had their geometry defined by total geometries it is possible to justify the formation of clusters from a conformational perspective. When the fragments are described with all distances or all angles there is remarkably good agreement between the clusters when the fragments were defined by these definitions and total geometries.

There are differences between distances only and angle only and on balance it appears that total geometries represent the best geometric definition of the fragments despite the high levels of redundancy generated by the definition. The use of distances only should not be discounted as there may be specific application, such as very large fragments, where the computational cost of clustering would be excessive. Using distances only in this context could allow a quicker but potentially less accurate clusters to be calculated.

Chapter 3

Triangles

3.1 Introduction

In this chapter there is an attempt to reduce the variables describing the conformation of the fragments. The common themes in this research is reducing the number of variables required to robustly and accurately describe the conformation of a fragment under investigation. As seen in the previous chapter, when a fragment has its geometry defined using total geometries there is a large degree of redundancy present. In essence, it is hoped that the redundancy of the geometric definition will be removed or at least reduced by combining the different variables together to form shapes. It is postulated that by measuring the area of these shapes it should be possible to summarise the variables that describes the shape. This section examines the nature of the variables that describe the fragments, and proposes a method of reducing the number of variables in a logical manner.

3.1.1 Different types of variables

During the examination of the variables in Chapter 2, it became apparent that not every variable was contributing an equal amount to the formation of clusters. The fragment 3-aminobutan-2-ol that was examined in the previous chapter will be used to illustrate this property. This fragment has two major conformational changes that give rise to the distribution of fragments in these

data. The dendrogram in Figure 2.11 gives an overview of the conformation of this fragment. There is also a minor conformational change where the carbon backbone is constrained by the chemical context from which the fragment was derived. When this fragment was clustered, the analysis yielded discrete clusters of fragments with similar conformations. The range of conformations within these data should allow the variables describing these fragments to be characterised. Table B.1 on page 141 shows descriptive statistics of each of the variables along with a figure indicating what part of the fragment the variables are describing. A smaller subset of this table is shown in Table 3.1. These tables show some basic statistics of these data describing the fragment. These include the range of the variable along with the maximum and minimum value of that particular variable. This gives an indication of the spread of each variable. There is also the standard deviation and the mean of the variable. The standard deviation gives an indication of the spread of that variable. This measure should only be trusted if the variable has a normal distribution. Examining the distribution of variables within these and other data it is apparent that normally distributed variables are an exception to what can typically be expected from these data. Typically a variable will have a bimodal or multi-modal distribution because the data are made up of fragments in discrete conformations. This is expected as the conformation of a fragment will typically fall into a global or local minimum on the energy hypersurface [5].

Using the sub sample in Table 3.1 as a representative sample of the variables describing this fragment, it appears that there are three different types of variable. Distance between atoms 1 and 2 (d_1_2) and the angle between atoms 2, 3 and 4 (a_2_3_4) are both bonded variables. As can be seen both the standard deviation (σ) and the range of the variables d_1_2 and a_2_3_4 are extremely small. It should be noted that these variables are not normalised and as a result angles will show more variability than distances. In contrast to the bonded variables, the remaining variables are measuring the relative distance between atoms which are not directly bonded or the angle between three atoms that are not joined directly by bonds.

The non bonded variables tend to vary more than the bonded variables.









Variable	σ	Range	Minimum	Maximum	Mean	Figure
d_1_2	0.019	0.116	1.447	1.563	1.524	
d_1_3	0.055	0.252	2.371	2.623	2.508	
d_1_4	0.417	1.566	2.340	3.906	3.056	
d_1_5	0.039	0.181	2.382	2.563	2.467	
a_1_3_4	22.223	79.401	69.454	148.855	98.090	
a_1_4_3	14.986974	54.308	19.368	73.676	53.536	
a_2_3_4	3.873	18.098	100.139	118.237	110.781	
a_1_4_6	15.1716	55.011	44.547	99.558	72.534	

Table 3.1: A selection of variables that describe the conformation of the fragment 3-aminobutan-2-ol. These variables were chosen to illustrate the how the distribution of each variable is effected by the chemical context from which the variables were derived.

Of course this is a rule of thumb and there are exceptions to this. An example of this is the distance between atoms 1 and 3 (d_1_3). When the distances are examined there is only a single distance that varies in a marked way within these data. The remaining three variables describing distances have an extremely small standard deviation. This is a consequence of the fragment from which these data were derived having a single conformational change.

Examining the angles in Table 3.1 it appears that there are three different types of angles within these data. There are the bonded angles that vary little (a_2_3_4). The variable with the highest degree of variability is the angle between atom 1, 3 and 4 (a_1_3_4) while the remaining variables lie somewhere in the middle. These 4 different types of variable are reproduced throughout these data. This can be seen in Table B.1 on page 141.

3.1.2 Semibonded angles

The variables which have a high variability have something in common. An example of a variable that has a high variability is the angle between atoms 1,3 and 4 and the statistics describing this variables can be found in Table 3.1. What this angle and the other angles that have a highest degree of variability have in common is that one of the rays of the angle is a bonded distance. A ray of an angle being one of the sides if the angle between the central atom and one of the outer atoms thus defining one half of an angle between three atoms. This is true for angle a_1_3_4 but is also true for angle a_1_4_3. Both of these variables have a common ray that is the bond between atoms 3 and 4. What is of interest is that both of these angles describe the same part of the fragment. Also it should be noted that when combined, these two angles form a triangle between the three points.

3.2 Triangles as a means of reducing variables

It was hoped that by applying this feature the number of variables necessary to describe the shape of fragment under investigation will be lowered. It is hoped that by calculating the area of a triangle between three atoms it will be possible to reduce the six variables describing the relative orientation of those atoms down to a single variable. In order to understand how the area of a triangle will vary in a molecular context, a simulation was carried out.

Initially there was an attempt to simulate the rotation of four atoms around a single bond. A schematic of the hypothetical molecule can be found in Figure 3.1.

3.2.1 Calculating the area of triangles

The calculation of the area of triangles was carried out using Heron's formula (Formula 3.2)[19, 58]. This formula calculates the area of a triangle using the lengths of the sides of the triangle. These lengths are taken from Figure 3.1. In this simulation, the bond lengths were fixed and the simulated molecule was rotated around the second bond to simulate a torsion.

$$s = \frac{1}{2}(a + b + c) \quad (3.1)$$

$$\Delta = \sqrt{s(s - a)(s - x)(s - y)} \quad (3.2)$$

where Δ is the area of the triangle being calculated.

In this case the lengths of x and y were calculated for various rotations around the bond b . These values are calculated in order to measure the area of the green triangle in Figure 3.1 x was calculated using Equation 3.3.

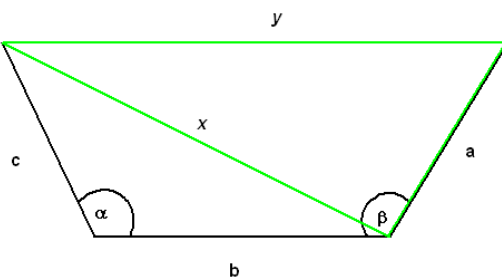


Figure 3.1: Sketch of the geometry of a hypothetical molecule. The triangle illustrated can be modified by altering any of the variables illustrated.

$$x = \sqrt{(c^2 + b^2) - 2bc(\cos \alpha)} \quad (3.3)$$

The calculation of y proved to be more troublesome. In order to simulate the rotation of a dihedral angle around b in Figure 3.1 it is necessary to treat the problem in an abstract manner. When there is a full 360° rotation around the central bond of a four atom system, the ‘shape’ of this system is shown in Figure 3.2. The shape is a conical frustum. To calculate the value of y for any value of α , β or τ the method in 3.4 - 3.12 were used.

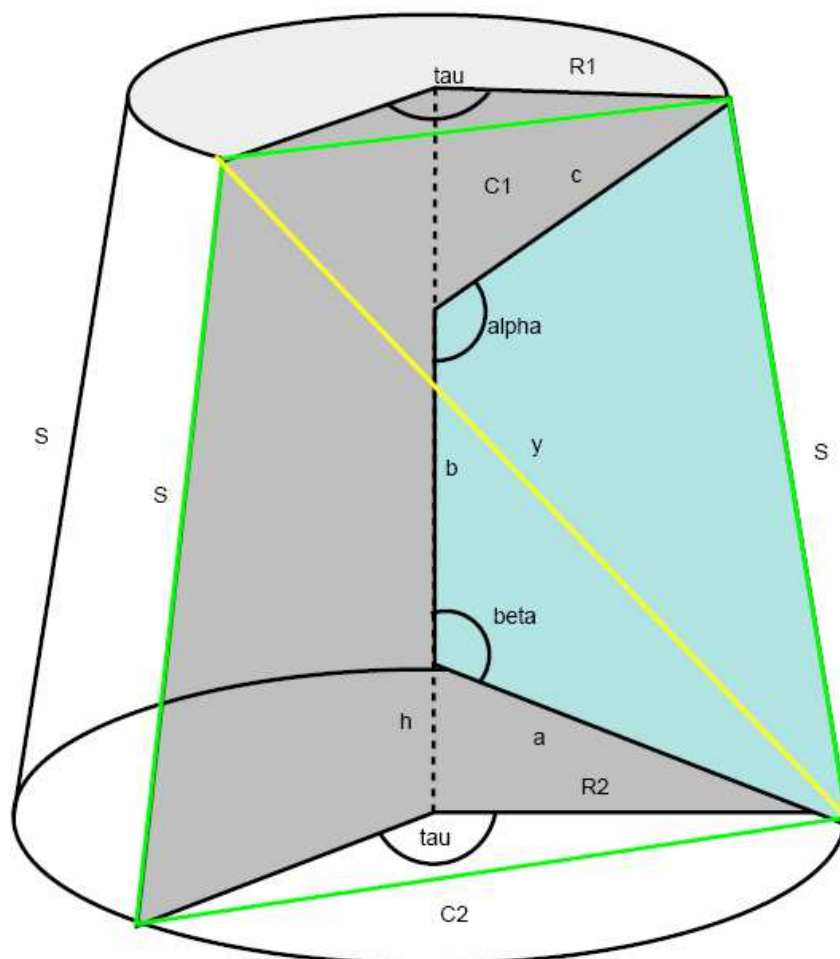


Figure 3.2: Diagram of simulated dihedral angle. From a geometric perspective the shape created by a rotation around a dihedral angle is a conical frustum. As the variables are changed the shape of the conical frustum will change. Using the method in Equations 3.4 - 3.12 the variables necessary for the calculation of an area of a triangle at any torsion angles can be calculated.

$$R_1 = \sin(\beta - 180)c \quad (3.4)$$

$$R_2 = \sin(\alpha - 180)a \quad (3.5)$$

$$h = b + (\cos(\alpha - 180)a) + (\cos(\beta - 180)c) \quad (3.6)$$

$$S = \sqrt{(R_1 - R_2)^2 + h^2} \quad (3.7)$$

$$C_1 = 2R_1 \sin\left(\frac{1}{2}\tau\right) \quad (3.8)$$

$$C_2 = 2R_2 \sin\left(\frac{1}{2}\tau\right) \quad (3.9)$$

$$y = \sqrt{\frac{C_1 C_2^2 - C_1^2 C_2 - C_1 S^2 + C_2 S^2}{C_2 - C_1}} \quad (3.10)$$

$$y = \sqrt{\frac{C_1 C_2 (C_2 - C_1) + S^2 (C_2 - C_1)}{(C_2 - C_1)}} \quad (3.11)$$

$$y = \sqrt{C_1 C_2 + S^2} \quad (3.12)$$

3.2.2 Area of triangles

This section shows the results of a simulation of a full 360° rotation around the central torsion with a range of α of between 100° and 120°. The β angle was constrained at 109° and the bond lengths were held at a length of two. The area of the triangle shown in Figure 3.1 was calculated for each of the increments and the results are displayed in Figure 3.3. As can be seen from the figure, the area of the triangle varies with the rotation around the central bond and for any given value of torsion, as the value of α increases the area of the triangle increases. This is no surprise given that the only distance that should vary with a rotational change is the distance y in this simulation. This distance is equivalent to d.1.4 in Table 3.1 and is the only variable describing a distance that significantly varies within this subset of the variables describing the conformation of the fragment 3-aminobutan-2-ol.

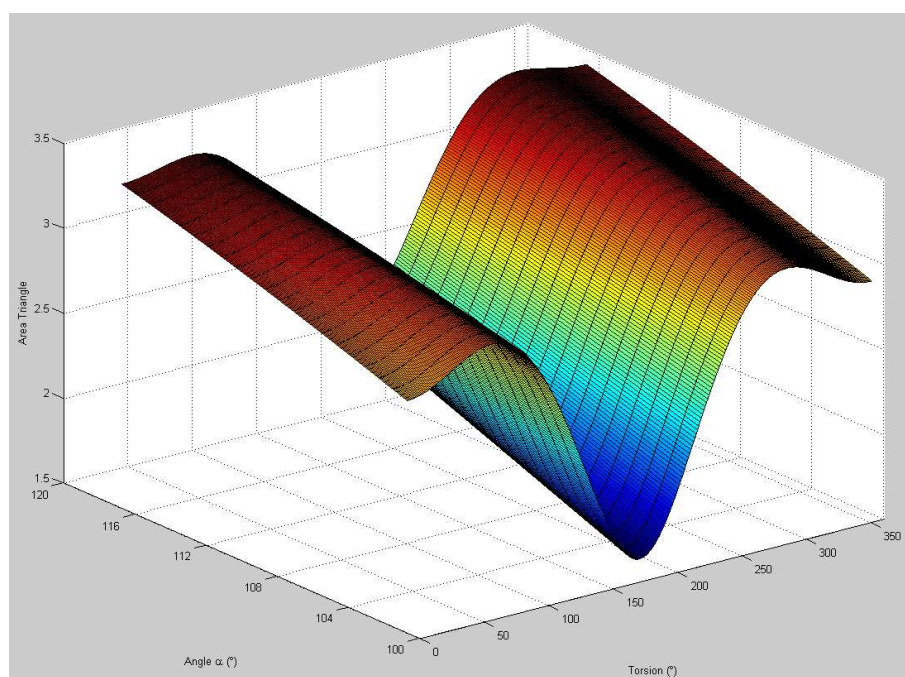


Figure 3.3: Graph of triangle area calculated for a range of torsion angles and a single bonded angle. The area of the triangle is plotted against the change of torsion angle and bonded angle α as illustrated in Figure 3.1 on page 72. The figure shows the area of a triangle changing over a range of $100^\circ - 120^\circ$ and a range of torsion of $0^\circ - 360^\circ$ ($-180^\circ - 180^\circ$)

3.2.3 Two dihedral angles with five atoms

The next simulation is aimed to simulate the area of a triangle affected by two independent torsion angles. In order to achieve this, a five atom fragment will have to be simulated. The triangle that will have its area measured has its apices at the atoms 1, 4 and 5. Once the coordinates describing the simulated fragment have been modified to simulate the position of the two torsion angles the distances between these coordinates are calculated using Pythagoras and the area of the triangle is calculated using Heron's formula [3.2]. The method by which the coordinates are modified can be found below.

3.2.4 Modifying coordinates to simulate a torsion

Initially the origin of the coordinate system is moved to the atom at the beginning of the of the torsion bond, i.e. the second atom of the four atoms in a torsion angle. This is illustrated in Figure 3.4 where the blue atoms indicated are placed at the origin of the coordinates.

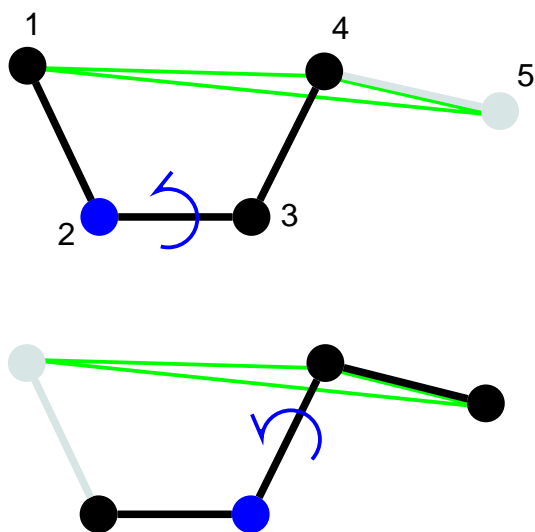


Figure 3.4: An illustration of the two torsion angles that are calculated with the triangle indicated. During the calculation the two torsion angles are calculated independently. This process involves aligning the simulated fragment on the blue atom and then modifying the coordinates of the atoms to the right of the blue atom. Once the appropriate coordinates have been modified the area of the triangle indicated in the figure is calculated.

$$\begin{bmatrix} x' \\ y' \\ z' \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & -p \\ 0 & 1 & 0 & -q \\ 0 & 0 & 1 & -r \\ 0 & 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} \quad (3.13)$$

Where p, q & r are the x, y & z coordinates of the atom that is to be moved to the origin.

The next phase is to rotate the axis of the system so that the central bond

lies on the x -axis. This involves rotating all of the coordinates around the y -axis [Equations 3.14] until the third atom of the torsion angle is in the yz plane of the coordinate system. At this point all the coordinates are rotated around the z -axis [Equations 3.16] until the third atom is on the x -axis.

Once the bond lies on the x -axis, all that is necessary to simulate the rotation around a torsion is to rotate the coordinates of the atoms that are further along the chain of atoms than the bond. In Figure 3.4, the atoms that have their coordinates modified are to the right of the blue atom. This is achieved using Equation 3.20 and simulates the increase in torsion angle for a single torsion.

Using this method it should be possible to recreate a system with 2 torsions by altering the atom at the beginning of the torsion angle. A schematic of the program is shown in Figure 3.5 and the output of this program is shown in Figure 3.6.

$$\begin{bmatrix} x' \\ y' \\ z' \\ 1 \end{bmatrix} = \begin{bmatrix} \cos \alpha & 0 & -\sin \alpha & 0 \\ 0 & 1 & 0 & 0 \\ \sin \alpha & 0 & \cos \alpha & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} \quad (3.14)$$

$$\begin{aligned} x' &= x \cos \alpha - z \sin \alpha \\ y' &= y \\ z' &= x \sin \alpha + z \cos \alpha \end{aligned} \quad (3.15)$$

Where α is the angle of rotation around the y axis.

$$\begin{bmatrix} x' \\ y' \\ z' \\ 1 \end{bmatrix} = \begin{bmatrix} \cos \beta & \sin \beta & 0 & 0 \\ -\sin \beta & \cos \beta & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} \quad (3.16)$$

$$x' = x \cos \beta + y \sin \beta \quad (3.17)$$

$$y' = -x \sin \beta + y \cos \beta \quad (3.18)$$

$$z' = z \quad (3.19)$$

Where β is the angle of rotation around the z axis.

$$\begin{bmatrix} x' \\ y' \\ z' \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos \gamma & \sin \gamma & 0 \\ 0 & -\sin \gamma & \cos \gamma & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} \quad (3.20)$$

$$x' = x$$

$$y' = y \cos \gamma + z \sin \gamma$$

$$z' = -y \sin \gamma + z \cos \gamma \quad (3.21)$$

Where γ is the angle of rotation around the x axis.

When the results of the simulation are examined in Figure 3.6 it appears that there is a large range of possible areas of triangles. This result is not entirely unexpected. It should also be noted that the smallest area of a triangle is close to zero. More pertinently, it should be noted that for every given area of triangle there could be a number of possible values for the two torsion angles. This poses a problem for using triangles as a measure of the conformation of two torsion angles. A given area of a triangle does not uniquely describe the conformation of the two torsions. This may prove to be problematic when fragments are described in *dSNAP* using triangles. While this is unfortunate it is not entirely unexpected and the use of the area of triangles as an input for *dSNAP* may still lead to a situation where the number of variables necessary to describe the conformation of a fragment is greatly reduced.

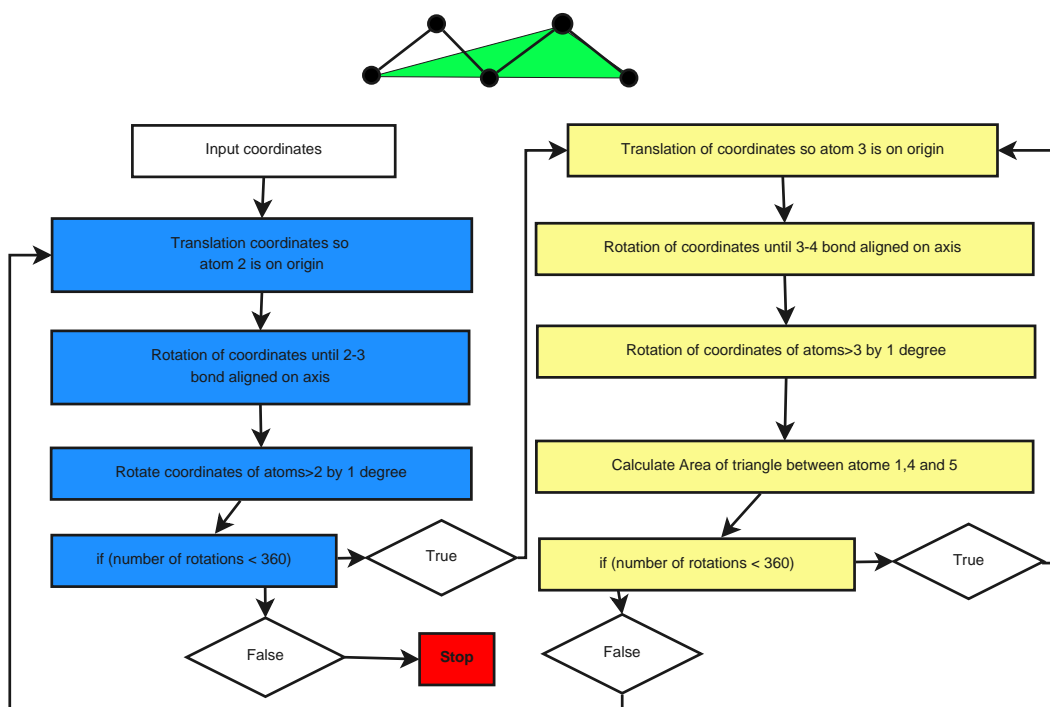


Figure 3.5: Description of the program that simulates a five atom fragment being rotated around two torsion angles. The area of the triangle being calculated is illustrated above.

3.2.5 Use as input to *dSNAP*

As an illustration of how triangles can be used to describe the geometry of a fragment the geometry of the fragment propan-2-one was described using triangles. This fragment has been previously analysed in Section 2.2.3. The dendrogram in Figure 2.19 and accompanying Newman projections give an illustration of the conformations that the fragment pentan-2-one takes in this example dataset.

Figure 3.8 shows the dendrogram, MMDS plot and cell display generated by *dSNAP* where the fragment propan-2-one was defined using the triangles illustrated in Figure 3.7. The area of these triangles was calculated using the orthogonal coordinates of the atoms as an input, Pythagoras to calculate the distances between the atoms and Heron's formula to calculate the area of these triangles. A list of the orthogonal coordinates was extracted from the CSD search and using a simple FORTRAN program the three atoms at

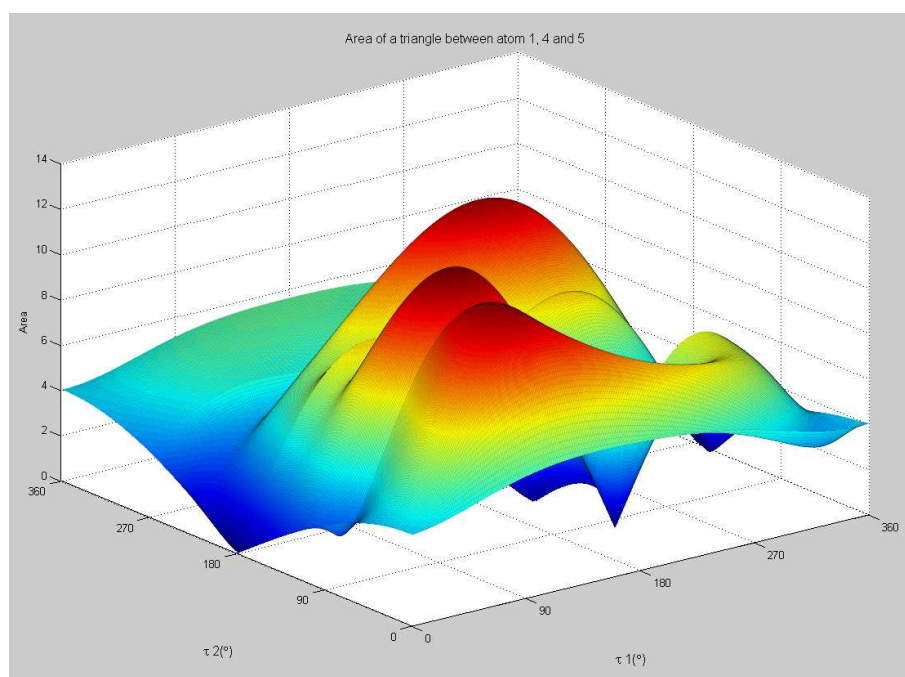


Figure 3.6: A surface plot representing the area of a triangle between atoms 1 4 and 5. An illustration of the fragment and the triangle being calculated can be found at the top of Figure 3.5. Torsion one (τ_1) is plotted against torsion two (τ_2) both of which range between 0° and 360°

the apexes of the triangle to be calculated were selected. The program then calculated the area of that triangle for all of the fragments in the search. This process was repeated for all of the triangles in Figure 3.7. The area of all of the triangles for each of the fragments were then tabulated and used as input for *dSNAP*.

When the fragments are examined in the fragment viewer, it appears that most of the fragments are in clusters with similar conformations. This is shown in Figure 3.9. While most of the fragments have been grouped into clusters with similar conformations, there are a number of fragments which have been grouped into a cluster where the conformation of these particular fragments vary a great deal from the average conformation within this cluster. This indicates that using triangles as a measure of geometry has not faithfully reproduced the clusters illustrated in Figure 2.19 where the fragment was defined by total geometries. By examining the MMDS plots in Figure

3.8, it is apparent that the discrete nature of the clustering achieved when the fragment was described using total geometries has been lost. In the right hand MMDS plot where that fragment was defined using triangles the fragments are more diffuse in appearance than when the fragment were defined with total geometries. This is an indication that this definition has failed to accurately separate fragments of clusters with similar geometry. Nevertheless, the number of variables required to make this approximation is vastly less than the number of variables that were used when the fragment were defined by total geometries. The results of clustering using this definition are discussed in Section 2.2.3. This section shows that non-bonded interactions are the most significant with regard to forming clusters of fragments with similar conformation. Using triangles as a measure of conformation has reduced the number of variables to a lower level than any of the definition of described in Section 2.2.3 but has unfortunately has not been successful in replicating the clusters formed in the on the left of Figure 3.8 where the fragment was defined with total geometries.

3.3 Conclusions

Triangles are an interesting method to summarise the shape of the fragments under investigation. This section has shown that it is possible to represent the rotation of a chain of atoms using the area of triangles as an indication of the conformation of the fragments. When a single torsion angle is simulated the resulting triangle is fairly simple to understand. The torsion angle simulated in Figure 3.3 shows as the torsion approaches 180° the area of the triangle being simulated reaches its minimum area. While this pattern is very easy to understand it could be regarded as somewhat overcomplicated. As noted previously there is only a single distance that changes in an idealised torsion angle. It should be noted that, excluding variation in bonded variables, there is little extra information gathered from calculating the area of a triangle across a torsion angle.

When two torsions are simulated the area of a simulated triangle across these torsions varies in a much less obvious manner. With reference to the

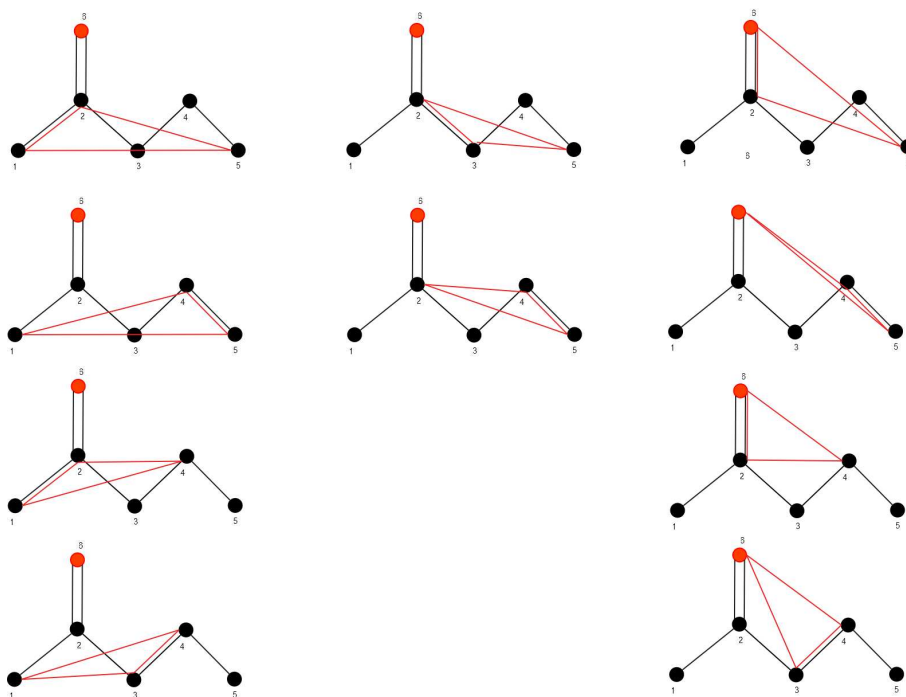


Figure 3.7: An illustration of the triangles used to describe the geometry of the fragment pentan-2-one. These 10 triangles were selected to describe all of the possible conformations of the fragment.

diagram in Figure 3.5, it appears that an area of a triangle is not unique to a given conformation of torsion angles. That is, for a given area of triangle there are a number of possible conformations for the two torsion angles. This poses a problem when using this definition to describe the geometry of a fragment. This could be one of the reasons why the fragment pentan-2-one was not grouped into clusters with similar conformations as was the case when the fragment was defined with total geometries.

It is possible that triangles could be used as extra variables during analysis with *d*SNAP. That is to say that triangles could be used in combination with other variables that combined will accurately describe the conformation of the fragment. This is an interesting proposal but there are a number of issues that should be addressed first. The fragment chosen was selected as it was close to a long chain of atoms which lends itself well to be described using triangles. This is because the major conformational changes are rotational

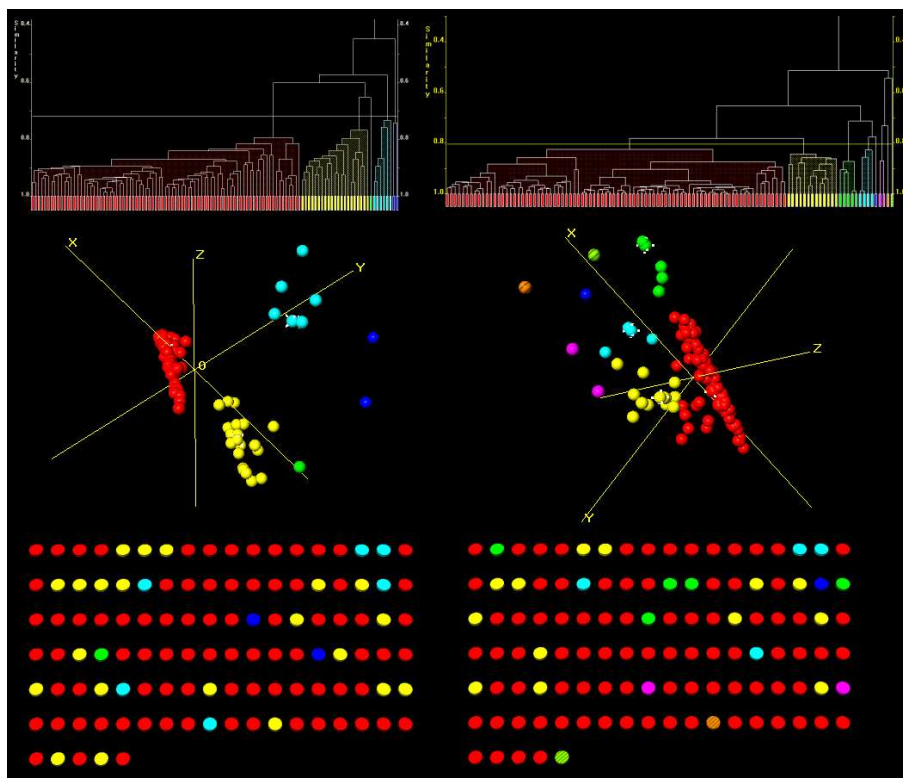


Figure 3.8: Combined MMDS plot, cell display and dendrogram of the fragment pentan-2-one where the fragment was defined with total geometries on the left and triangles on the right.

in nature. Detecting differences in rotation could be measured using torsion angles but as discussed in Section 2.2.3.5 there currently is a problem using torsion angles in *d*SNAP. This is the result of the rotational nature of the torsion angles where in absolute terms the difference between torsion angles can be large. For example, the difference between -179° and 170° is 358° if the rotational nature of torsion angles are ignored. In practice, the difference between the two torsion angles above will be 2° . There are also a number of situations where triangles may not be ideally suited to describing the conformation of fragments. In the stated example the choice of triangle is relatively obvious. If, for instance, a fragment that was centred on a metal atom, triangles would not be ideally suited to describing the conformation of this fragment. There is also the problem that the description of the fragment

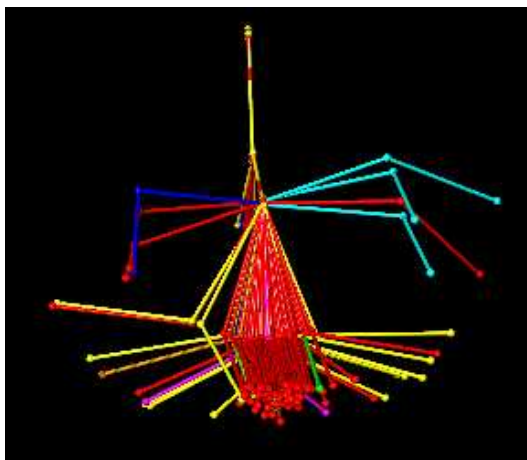


Figure 3.9: Fragment view of pentan-2-one where the fragment was defined with triangles. In this figure the fragments have been aligned in such a way that it is obvious that some of the fragments have been grouped into clusters that are not of broadly single conformation.

by triangles is more abstract than with total geometries. As a result it is more difficult to justify changes in conformation of the fragment based on the variables. One of the key advantages of using total geometries to describe the geometry of the fragments is that variables can be referred to justify the formation of clusters.

Chapter 4

Factor Analysis

4.1 Introduction

Factor analysis (FA) is a statistical method that attempts to find underlying trends in data sets. It has its origins in psychology and in essence has the aim of describing and examining the internal structure of a correlation or a covariance matrix. Generally, statistical tests are applied to study the relationship between independent and dependent variables. Factor analysis differs from these statistical tests. Factor analysis aims to discover underlying features or patterns of dependent variables with the goal of uncovering independent effects or influences on the datasets that are not directly measured. As a result, these factors are necessarily more hypothetical because these variables are not actually measured even if they could be. A typical use of factor analysis aims to uncover how many factors are required to explain the pattern and relationship between the variables in the data. The analysis also aims to give some meaning to the nature of these factors along with measures of how well these factors explain the data. It is hypothesised that given the context under which the data are generated, then a factor could be attributed to a specific conformational change within the fragment under investigation.

The most prominent pioneers of factor analysis were Spearman[52], Thomson[55], Thurstone[56] and Burt[12]. Spearman set out to try and

find an underlying trait for general intelligence that he termed g . It was hoped that g could be characterised using easily measured variables, such as mathematical skill and verbal reasoning. The measurement of these variables can then be used to extrapolate the latent variable g . As the name suggests these latent variables are hypothetical variables that measure hidden features in the dataset. Spearman's research in this area was a continuation from previous research looking at the correlation between different traits with the aim of finding an underlying relationship between them. Spearman aimed to expand the scope of this research to find the more general characteristic of general intelligence [52]. Unfortunately this study came to no real conclusion in terms of uncovering a measure of general intelligence.

One method of performing factor analysis is to begin with principal component analysis and use the first few principal components as unrotated factors. This is a simple method and is a good starting point for factor analysis. With p variables there will be p principal components that are linear combinations of the original variables [42].

$$\begin{aligned} Z_1 &= b_{11}X_1 + b_{12}X_2 + \dots + b_{1p}X_p \\ Z_2 &= b_{21}X_1 + b_{22}X_2 + \dots + b_{2p}X_p \\ &\vdots \\ Z_p &= b_{p1}X_1 + b_{p2}X_2 + \dots + b_{pp}X_p \end{aligned}$$

Where b_{ij} values are given by the eigenvectors of the correlation matrix calculated from the original data and X are the p variables. The transformation from X values to Z values is orthogonal, so that the inverse relationship is

$$\begin{aligned}
X_1 &= b_{11}Z_1 + b_{21}Z_2 + \dots + b_{p1}Z_p \\
X_2 &= b_{12}Z_1 + b_{22}Z_2 + \dots + b_{p2}Z_p \\
&\vdots \\
X_p &= b_{1p}Z_1 + b_{2p}Z_2 + \dots + b_{pp}Z_p
\end{aligned}$$

For factor analysis, only m of the principal components are retained. The numbers of factors are chosen by the user. The value e_i is a linear combination of the Z_{m+1} to Z_p . All that remains to do is scale the principal components such that they have unit variance which is a requirement for factors. This is achieved by dividing Z_i by its standard deviation. In this case it is the square root of the i th eigenvalue of the correlation matrix ($\sqrt{\lambda_i}$). The equations now become

$$\begin{aligned}
X_1 &= \sqrt{\lambda_1}b_{11}F_1 + \sqrt{\lambda_2}b_{21}F_2 + \dots + \sqrt{\lambda_m}b_{m1}F_m + e_1 \\
X_2 &= \sqrt{\lambda_1}b_{12}F_1 + \sqrt{\lambda_2}b_{22}F_2 + \dots + \sqrt{\lambda_m}b_{m2}F_m + e_2 \\
&\vdots \\
X_p &= \sqrt{\lambda_1}b_{1p}F_1 + \sqrt{\lambda_2}b_{2p}F_2 + \dots + \sqrt{\lambda_m}b_{mp}F_m + e_p
\end{aligned}$$

where $F_i = Z_i/\sqrt{\lambda_i}$.

The unrotated factor model is then

$$\begin{aligned}
X_1 &= a_{11}F_1 + a_{12}F_2 + \dots + a_{1m}F_m + e_1 \\
X_2 &= a_{21}F_1 + a_{22}F_2 + \dots + a_{2m}F_m + e_2 \\
&\vdots \\
X_p &= a_{p1}F_1 + a_{p2}F_2 + \dots + a_{pm}F_m + e_p
\end{aligned}$$

Where $a_{ij} = \sqrt{\lambda_i}b_{ij}$ [42].

After rotation, which in this case was Varimax [37], the new solution is

$$\begin{aligned} X_1 &= g_{11}F_1^* + g_{12}F_2^* + \dots + g_{1m}F_m^* + e_1 \\ X_2 &= g_{21}F_1^* + g_{22}F_2^* + \dots + g_{2m}F_m^* + e_2 \\ &\quad \vdots \\ X_p &= g_{p1}F_1^* + g_{p2}F_2^* + \dots + g_{pm}F_m^* + e_p \end{aligned}$$

where F_i^* is the new i th factor after rotation. The rotation takes place to make the interpretation of the factors easier.

Varimax rotation is based on the assumption that the interpretability of a factor j can be measured by the variance of the square of its factor loadings. That is the variance of $a_{1j}^2, a_{2j}^2, \dots, a_{mj}^2$. If this variance is large, the value of a_{ij} tend to be either large or close to zero. Varimax rotation therefore aims to maximise the sum of these variances for all of the factors.

The value of the i th unrotated factor is just the i th principal component that has been scaled to have unit variance. The values of the rotated factors are more difficult to obtain. These rotated factors can be calculated using the following formula:

$$\mathbf{F}^* = \mathbf{X}\mathbf{G}(\mathbf{G}'\mathbf{G})^{-1} \quad (4.1)$$

Where \mathbf{F}^* is an $(n \times m)$ matrix of the values for the m rotated factors and n original rows of data. \mathbf{X} is the $(n \times p)$ matrix of the original data of p variables and n observation that have been standardised to have a mean of zero and unit variance. \mathbf{G} is the $(p \times m)$ matrix of rotated factor loadings [42].

The most important outputs of factor analysis are a matrix of factor loadings and a column of communalities. Factor loadings represent the extent to which each of the variables is related to the hypothetical factor. In some methodologies of factor analysis the factor loadings can be regarded as correlations of these variables to the hypothetical factors. Communalities represent the sum of squares of the factor loadings. These communalities represent the extent of the overlap between variables. If the communalities

are 1.0 then the variance of that variable can be explained with the weighted combination of the factor loadings. If the communality is 0.0 then the variance of the variable does not share anything in common with the factors calculated.

Being able to understand what these factors are indicating is far more important. It is obvious that a variable with a high factor loading is a good indicator of what this factor is describing. It is equally informative if a variable has an extremely low loading. This of course leaves the middle ground and the difficult question of what is an important factor and what is not. Unfortunately there is no statistical test to give a clear indication of what is a significant factor loading when the loading matrix has been rotated. This is because the rotation can be regarded as arbitrary, or at least a means to an end. The rotation is carried out in order to simplify the interpretation of the analysis. As a result of this, the selection of a level of significance could be regarded as somewhat subjective. Comrey [16] notes that a common cut-off of significance for a factor loading is 0.3 where the factor loadings are orthogonal. This is because a factor loading of 0.30 when squared gives a value of 0.09. This means that a variable with a factor loading of less than 0.3 shares less than 10% of its variance with the hypothetical factor. Comrey and this author believe that this arbitrary cut-off is rather low, especially when the variables are highly correlated which is the case here. A factor loading of greater than 0.71 in this context seems more appropriate. This factor loading indicates that a variable with this loading shares 50% of its variance with the hypothetical factor. Of course, if a large number of the variables have an extremely high factor loading, then the threshold can be set much higher. This means that the number of variables that can be discarded can be increased while still maintaining the structure within the data matrix.

The purpose of factor analysis is to discover simple patterns in the relationships among a set of variables. In particular, it seeks to discover if the observed variables can be explained largely or entirely in terms of a much smaller number of variables called factors. Unlike many statistical methods which study the relation between independent and dependent variables factor analysis is used to study the patterns of relationships found among many de-

pendent variables. The goal of factor analysis is to discover something about the nature of the independent variables that affect the pattern of relationships despite the independent variables not being measured directly. As a result, answers derived using factor analyses are more hypothetical and tentative than if the independent variables are observed directly. These inferred independent variables are called factors. A typical factor analysis proposes answers to four major questions [41, 31]:

1. How many factors are needed to explain the pattern of relationship among these variables?
2. What is the nature of those factors?
3. How well do the hypothetical factors explain the observed data?
4. How much purely random or unique variance does each of the observed variables include?

4.2 Example: Finding Common Factors Affecting Exam Grades

This Example was adapted from [43]. 120 students have each taken five exams, the first two covering mathematics, the next two on literature, and a comprehensive fifth exam. It seems reasonable that the five grades for a given student ought to be related. Some students are good at both subjects, some are good at only one, *etc.* The goal of this analysis is to determine if there is quantitative evidence that the students' grades on the five different exams are largely determined by only two types of ability.

Factor analysis was applied to these variables and factor loadings for two factors were extracted. These were not rotated. From the table of factor loadings (4.1), you can see that the first unrotated factor puts approximately equal weight on all five variables, while the second factor contrasts the first two variables with the second two. You might interpret these factors as "overall ability" and "quantitative vs. qualitative ability" This also shows

Test	Factor 1	Factor 2
Math 1	0.6289	0.3485
Math 2	0.6992	0.3287
Literature 1	0.7785	-0.2069
Literature 2	0.7246	-0.2070
Comprehensive	0.8963	-0.0473

Table 4.1: Table of factor loadings for the model example

that the comprehensive test is the test that best represents the first factor and therefore “overall ability”

4.3 Application to *d*SNAP

It was hoped that by applying factor analysis to the data used in *d*SNAP it will be possible to reduce the number of variables required to describe the formation of each cluster and will allow easier interpretation of the reasons why clusters have formed. By applying factor analysis to total geometries, it should be possible to remove variables that have little or no contribution to the formation of clusters. It was also hoped that the latent underlying factors are actually conformational changes within these data and will aid the understanding of the formation of clusters. Factor analysis was carried out using the SPSS software package [54]. The factors were extracted using principal components analysis and varimax rotation was applied to simplify the process of analysing the factors. In this analysis six factors were initially extracted and using the cumulative variance of the rotated sum of squared loadings along with the table of factor loadings a threshold was chosen. The variables that had a loading greater than or equal to the threshold were extracted from the original data matrix and tabulated. The tabulated data was then used in *d*SNAP.

4.4 3-chlorobut-2-ene-thiolate

A description of the clustering of this fragment can be found in Section 2.2.1.

Factor analysis was applied to the variables describing the fragment 3-chlorobut-2-ene-thiolate and the number of variables was reduced from 75 variables to 46 variables. The variables extracted by this process were d23, d34,d36, d45, d56, a213, a314, a315, a316, a415, a416, a516, a123, a125, a324, a326, a425, a426, a526,a134,a135, a136, a234,a236, a435, a536, a143, a145, a243, a245, a246, a346, a546, a154, a156, a253, a254, a256, a354, a356, a163, a165, a263, a265, a364, a465. The variables were chosen by selecting those variables with a factor loading of greater than $|0.9|$. This level as chosen was a lower threshold would include many more variables which would render the application of factor analysis useless. When the factor loading for a given variable was greater than the threshold, this variable was deemed to be significant and the raw data for this particular variable was extracted. The collated significant variables were then used as the input for *d*SNAP. Examining the factor loadings in Table C.1 on page 147, all of the 46 variables were extracted from the first 3 factors. By examining Table C.3 on page 151 the cumulative percentage of variance explained was 83.3%. When these variables have been extracted and tabulated these data were used as input for *d*SNAP and the results are shown in Figures 4.1 and 4.2.

Figure 4.1 shows a cell display where the cluster that each fragment belongs to is represented by the colour of a circle. These colours are taken from the dendrograms in Figure 4.2. By examining the colours and therefore the cluster that each of the fragments are in, it is apparent that by reducing the number of variables with the application of factor analysis, the clusters have been preserved. When the dendrograms are examined in Figure 4.2, it shows that the clustering is close to identical between total geometries and when the number of variables has been reduced by the application of factor analysis. There are a few rearrangements within each cluster but this reflects minor perturbations in the distance matrix, not a major difference in the classification of the fragments. A problem with the application of factor analysis in this example is that the conformational change in this example is so simple and the geometric definition so highly correlated that the variables have extremely high loadings in the factor loadings. Table C.1 shows the tabulated rotated factor loadings. This could be the reason why there are so

many ‘important’ variables selected by the application of factor analysis. It may be possible to constrain the criterion that was used to select the variables further and as a result lower the number of variables selected while still preserving the clustering. Of course the application of a threshold to select variables is arbitrary and as a result there is no correct answer to where to draw the line where variables are ‘important’ or not.

4.5 3-aminobutan-2-ol

The analysis of the cluster of these data is described previously in Section 2.2.2.

When factor analysis was applied to these data, 19 variables were extracted from the 75 original variables as having a factor loading greater than $|0.9|$. Variables were extracted from the first 4 factors that describe 62.3% of the variance in these data (Table C.6). The variables extracted were: d16, d46, d56, a216, a316, a415, a126, a136, a436, a536, a346, a154, a254, a356, a162, a163, a265, a365. The procedure to select these important variables was exactly the same as the above section. The reduced data set was then run through *d*SNAP the results of clustering can be seen in figures 4.3 and 4.4. As can be seen in the Figure 4.3 there is good agreement between the 2 different definitions of the geometry of the fragments. Since 62.3% of the variance of these data is explained, it should not be expected that there will be exact agreement between the two different geometric definitions. Nevertheless the reduced data represents 20% of the original variables. In this case it should make the process of understanding what variables are causing the formation of clusters easier. In further chapters, visualisation of these variables will be shown in a biplot.

4.6 Pentan-2-one

The analysis of the cluster of these data is described previously in Section 2.2.3.

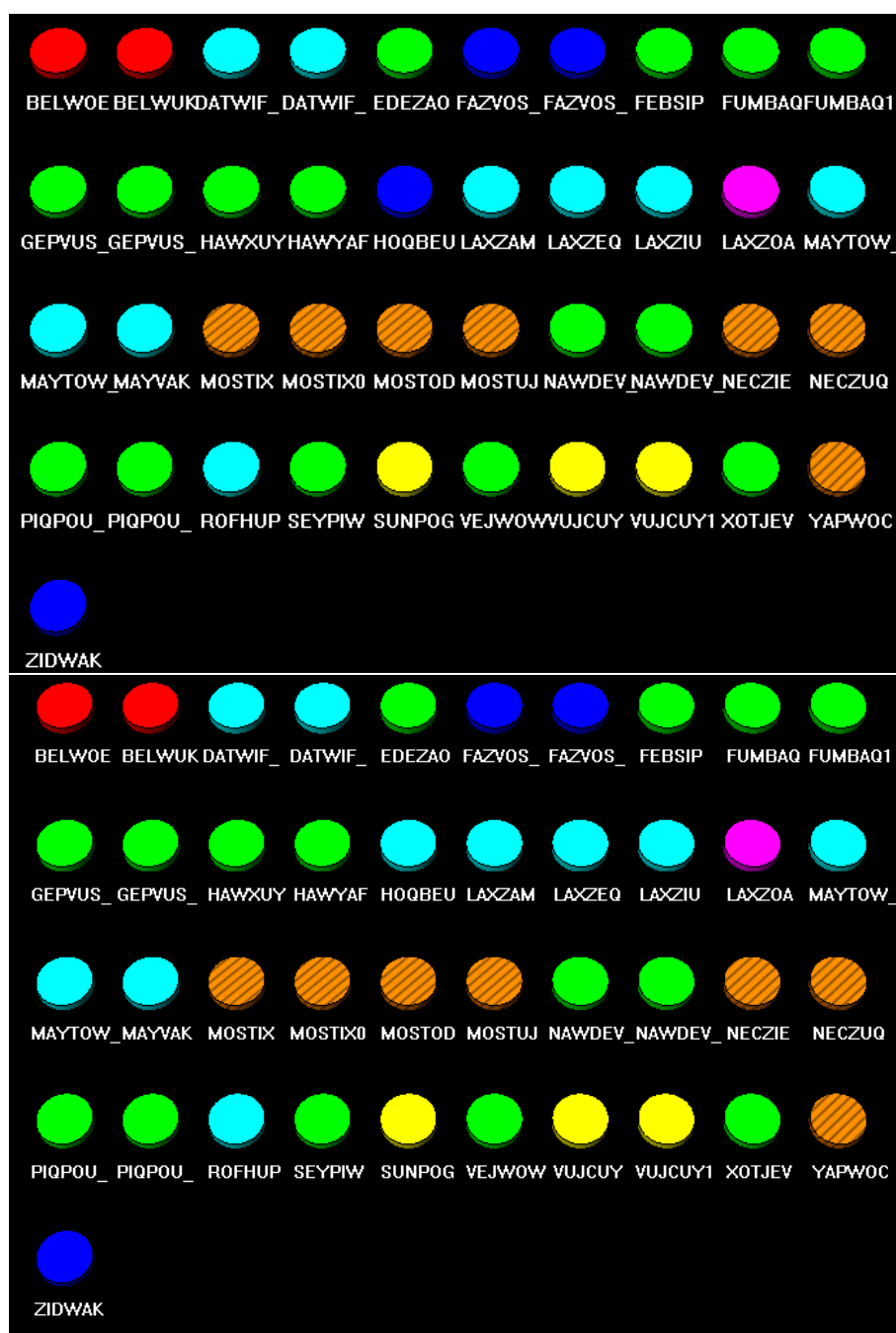


Figure 4.1: Cell displays of 3-chlorobut-2-ene-thiolate with the fragments defined with total geometries on the top and with the variables reduced by the application of factor analysis on the bottom.

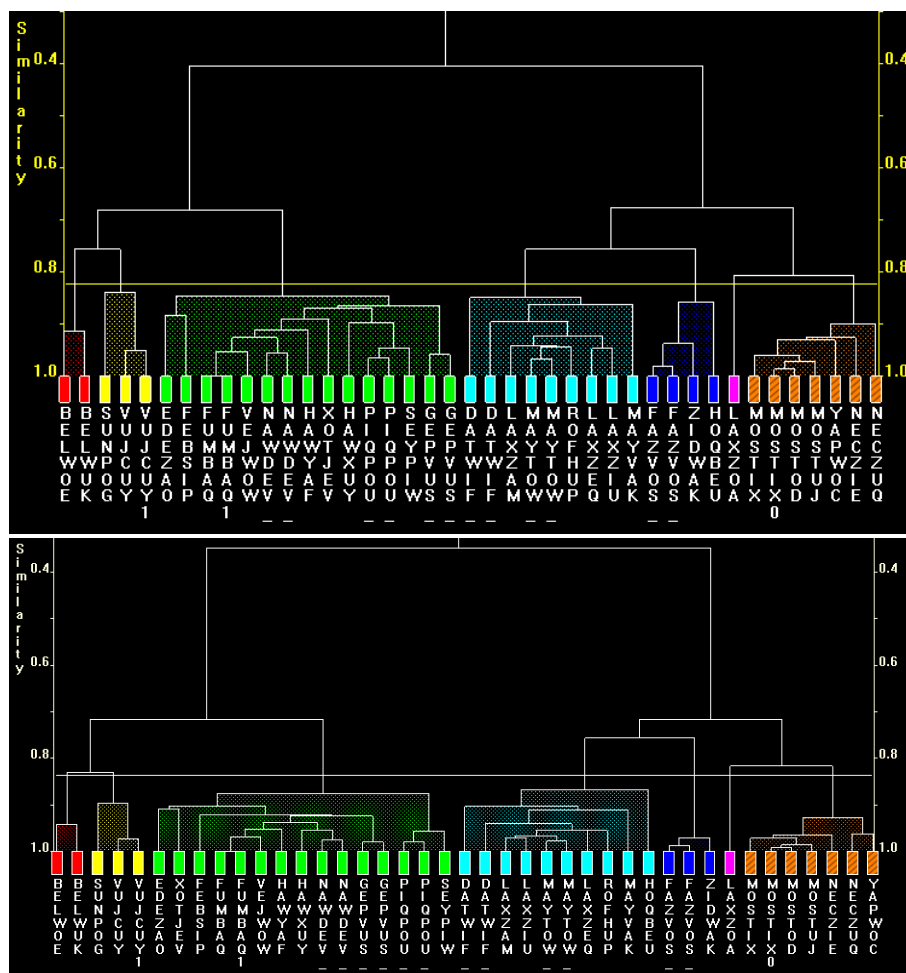


Figure 4.2: Dendrogram of 3-chlorobut-2-ene-thiolate with the fragments defined with total geometries on the top and with the variables reduced by the application of factor analysis on the bottom.

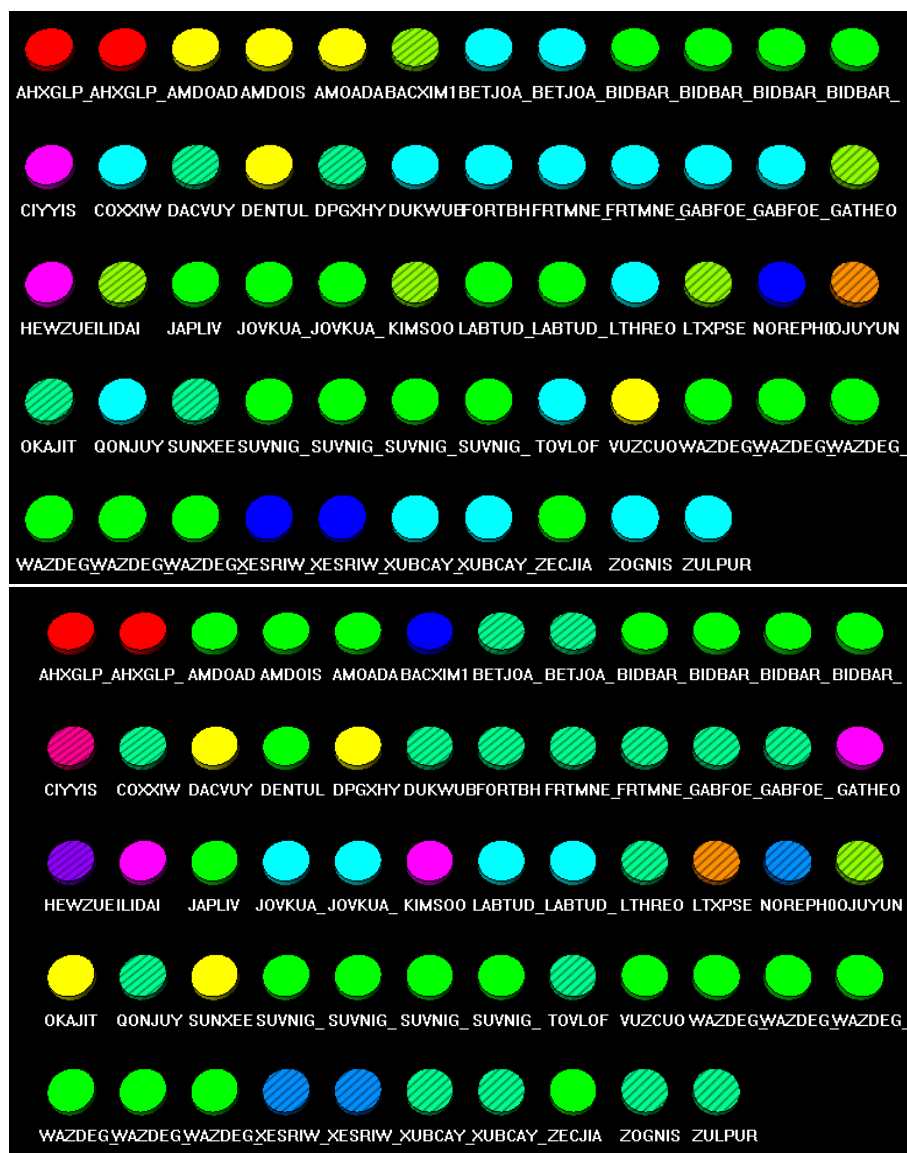


Figure 4.3: Cell display of the 3-aminobutan-2-ol with the geometry of the fragments defined by total geometries on the top and with the variables reduced by the application of factor analysis on the bottom.

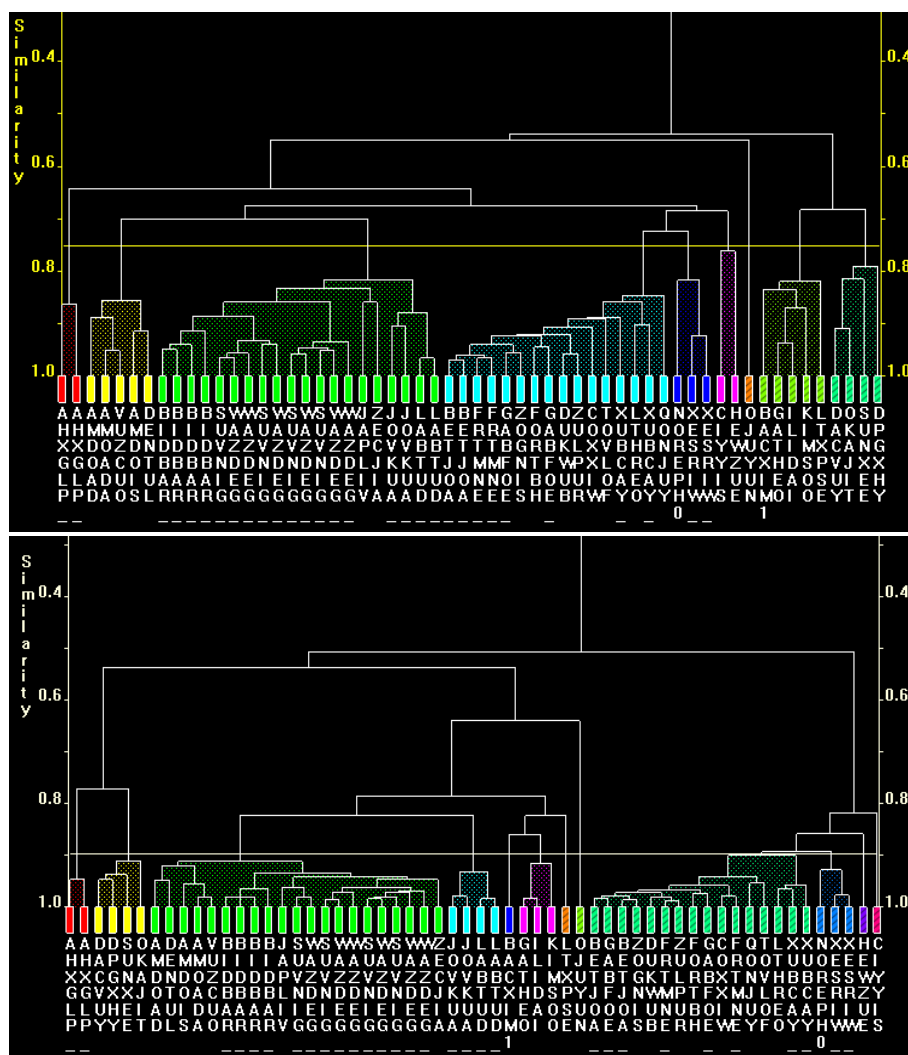


Figure 4.4: Dendrograms of the 3-aminobutan-2-ol with the geometry of the fragments defined by total geometries on the top and with the variables reduced by the application of factor analysis on the bottom.

When factor analysis was applied to these data, 40 variables were extracted as having a factor loading greater than $|0.9|$ from the original 75 variables. These were d14, d25, d46, d56, a214, a216, a316, a415, a416, a123, a124, a325, a425, a426, a132, a234, a235, a236, a435, a436, a536, a142, a145, a146, a245, a246, a345, a346, a546, a154, a253, a254, a256, a354, a356, a456, a164, a264, a365, a456. Variables were extracted from the first 5 factors that describe 84.46% of the variance in these data. These variables were then collated and analysed in *dSNAP*. The results from the clustering can be seen in Figures 4.5 and 4.6 It can be seen from the cell displays in Figure 4.5 that when the fragment was defined using the variables selected by factor analysis, the clustering was almost identical to the fragments defined by total geometries. There were only differences between how the fragments were related, not how they were grouped together. Since 85% of these data were explained by 5 factors it should be expected that there would have been good agreement between results of the analysis in *dSNAP*.

4.7 Conclusions

Factor analysis successfully reduced the number of variables required to describe the geometry of the fragments significantly. When these variables were extracted from the original data and analysed in *dSNAP* it appeared that these reduced data have, for the most part, generated clusters that are in good agreement with the clusters formed when the fragments were defined by total geometries. This indicates that these reduced data are approximating to a high degree of accuracy distance matrix generated by total geometries. Nevertheless it would be unwise to use this method as an initial treatment to datasets that are becoming unwieldy. The application of factor analysis should be reserved for the post cluster analysis interpretation of the results from *dSNAP*.

It is traditional that factors should be given names that describe the properties of the data that this factor describes. Typically, these factors are named after the variables that contribute towards the factor. For example, if 2 variables measuring the height and weight of a population are related to a

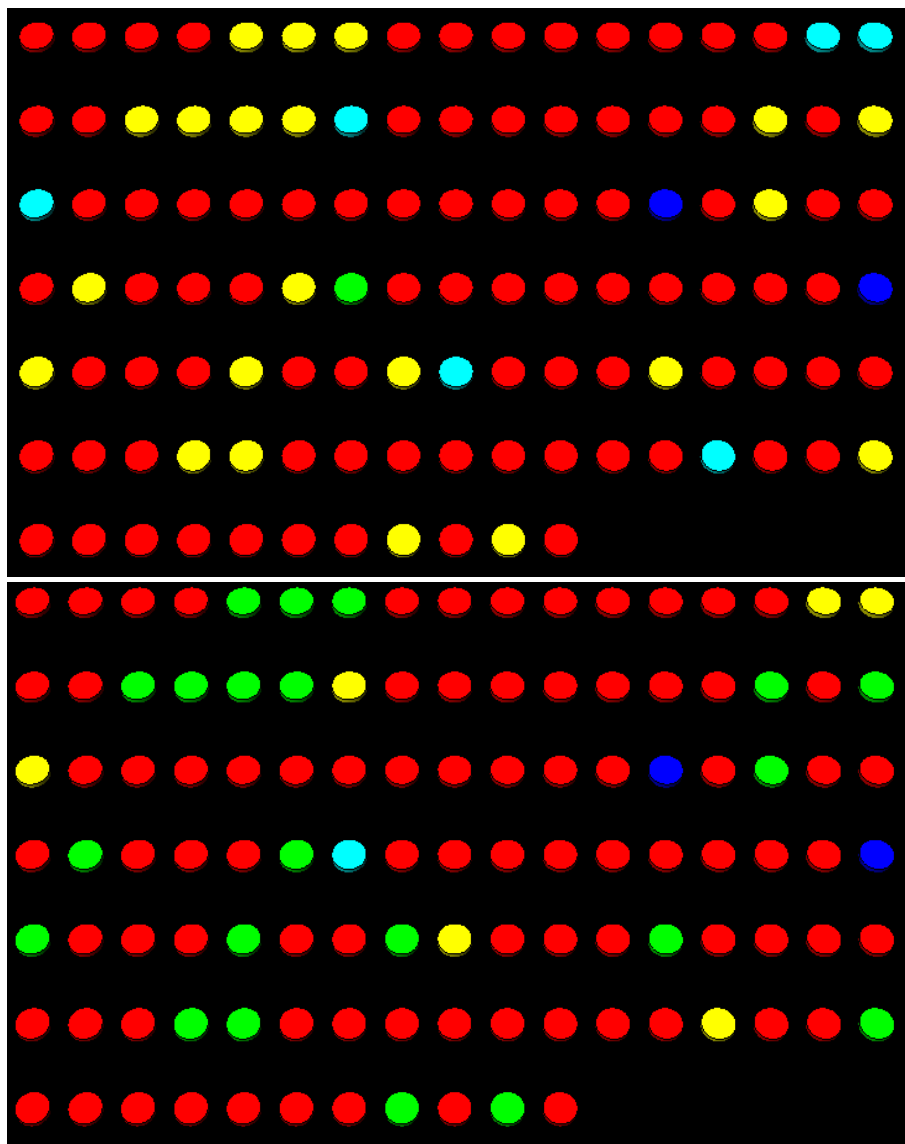


Figure 4.5: Cell display of the pentan-2-one with the geometry of the fragments defined by total geometries on the top and with the variables reduced by the application of factor analysis (bottom).

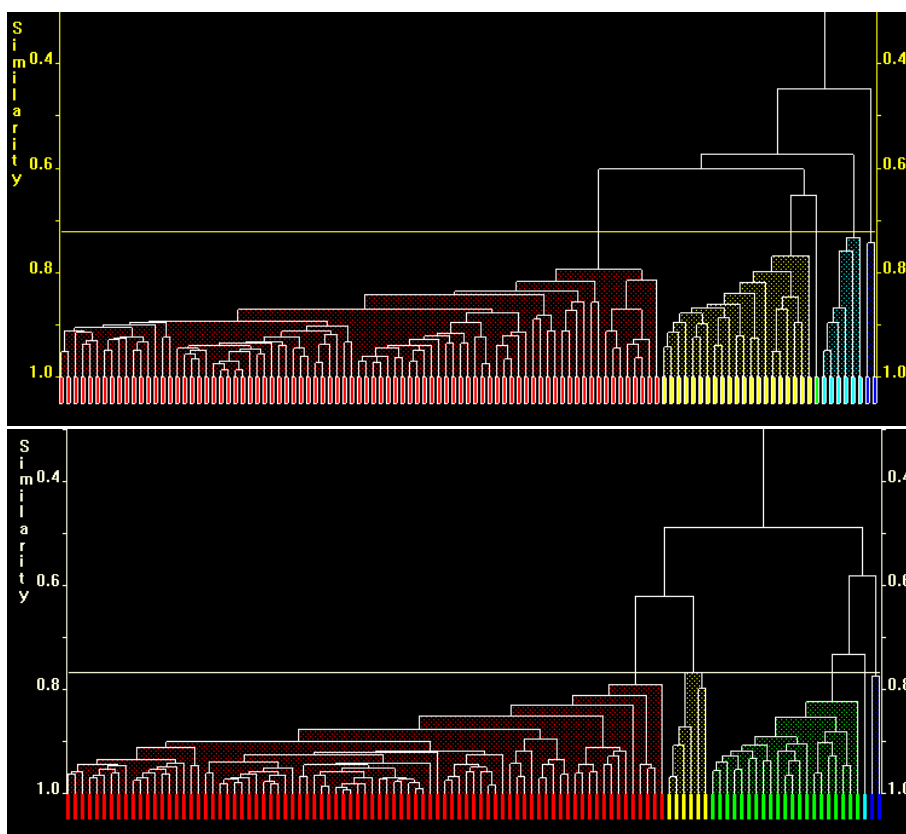


Figure 4.6: Dendrograms of the pentan-2-one with the geometry of the fragments defined by total geometries on the left and with the variables reduced by the application of factor analysis (bottom).

factor this factor could be named 'height and weight'. In this analysis all of the variables are highly interrelated. As described previously each variable does not describe a single change in conformation. Equally, a variable may describe more than 1 conformational change. In light of this, all attempts to assign names to factors extracted in this chapter have proven to be futile.

Using the reduced variables it is easier to detect trends in the data. It would appear from the clustering that the variables selected by the application of factor analysis represents variables that can describe the geometry of the fragments. It is hoped that these factors should represent the underlying latent variables that correspond to different conformational changes. The understanding of these factors in the context of conformational changes is difficult just from tabulated data. The next chapter explores a plotting method that may be used to visualise the factors with the aim of understanding which factor is representing what conformational change.

Chapter 5

Biplots

Biplots [28] [29] [25] are regarded as a multivariate analogue of scatter plots and were developed by K.R. Gabriel[25]. Biplots aim to create a plot that shows both the samples and the variables describing these samples in the same plot. In order to reduce the dimensions of the data set, principle components are used. The variables are represented by axes, where the cosine of the angle between axes approximates the correlation between these variables and the length of these axes approximates the standard deviation of that variable. These features are illustrated in Figure 5.1. Within this figure, the diagram on the left shows a biplot drawn for the purpose of illustration. The lengths of each coloured line represent the standard deviation of that variable. In Figure 5.2, the blue line in the diagram on the left is of higher standard deviation than the red and yellow lines. The angle between axes representing the variables is given by the *cosine* of the correlation between these variables. Thus, the biplot can give an easy to understand overview of a correlation matrix in both 2 and 3 dimensions as well as an indication of how variable the variables are.

This is only part of the story. The reason why a biplot is called a biplot is that all the samples and all the variables are plotted on a single plot. The process where the samples are plotted onto the biplot is called interpolation. This process is illustrated in Figure 5.1 where the samples are placed orthogonally to the approximate value of all the variables that describe this

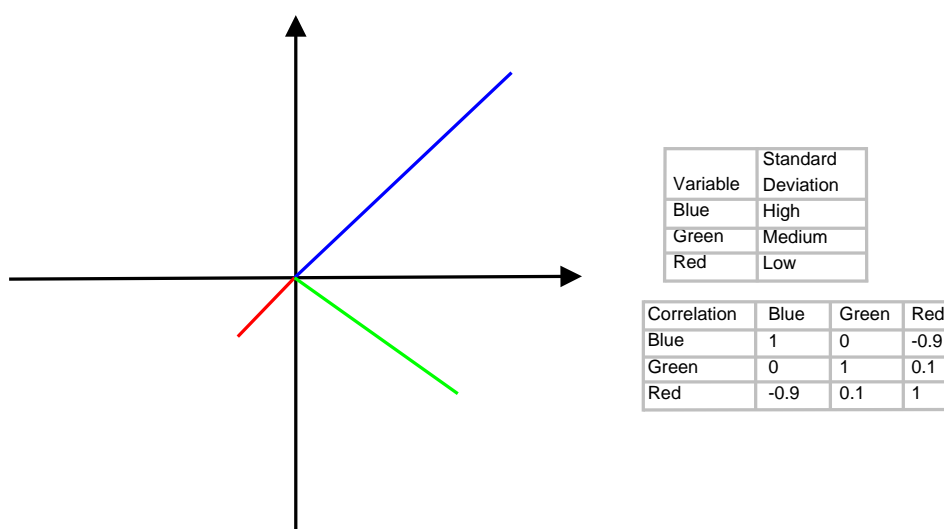


Figure 5.1: An illustrative biplot with the corresponding correlation matrix and an indication of the standard deviation. This biplot aims to illustrate the relationship between the variables in a biplot. The cosine of the angle between the variables represents the correlation between variables while the length of the axis represents the standard deviation of the variables.

fragment. These two features of biplots make them an extremely useful tool for examining the reasons for the formation of clusters in *d*SNAP.

In order to illustrate the properties of a biplot an imaginary dataset is displayed in Figure 5.3. This figure shows a data set of 15 people where their height, weight and hair length were measured for each individual. The plot on Figure 5.3 shows that there is a high correlation between height and weight while there is a low correlation between hair length and both height and weight. With regards to the samples, these were coloured according to sex and there are clear trends between the two populations. It is clear that the males tend to be taller, heavier with shorter hair while the females tend to be shorter, lighter with longer hair.

What is the value of biplots? A biplot allows a user to choose variables which may be significant to the formation of clusters. Once the nature of the plot is understood, it becomes apparent that a biplot is a useful tool to uncover the reasons behind the formation of clusters. This is illustrated in the following examples.

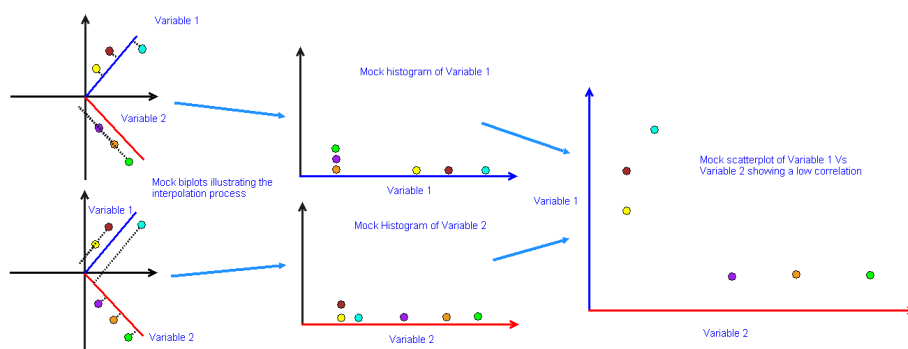


Figure 5.2: A figure illustrating the properties of the process of interpolation in the context of biplots. The interpolation of the samples in a biplot gives an indication of what value that sample, or fragment in this context, has for each of the variables describing the sample. The interpolation process aims to place a sample orthogonally from the values that the fragment has for that variable. The result of this process means that by examining the position of the sample relative to the variables, it is possible to understand what changes in variables have resulted in the fragment being placed where it is on the plot.

5.1 Calculating biplots

In this work, biplots were derived from principal components analysis resulting in a $n \times p$ matrix \mathbf{X} that gives the coordinates of n samples described by p variables. The objective of principal components analysis is to take the p variables X_1, X_2, \dots, X_p and find combinations of these variables to produce indices Z_1, Z_2, \dots, Z_p that are uncorrelated in order of their importance. These indices describe the variation in the data. Principal components analysis involves finding the eigenvalues of the sample covariance matrix. If the $n \times p$ data matrix has been standardised such that each variable has zero mean and unit variance, the matrix is a correlation matrix. The variances of the principal components are the eigenvalues of the covariance matrix. Assuming that the eigenvalues are ordered $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$, then λ_i corresponds to the i th principal component

$$Z_i = a_{i1}X_1 + a_{i2}X_2 + \dots + a_{ip}X_p \quad (5.1)$$

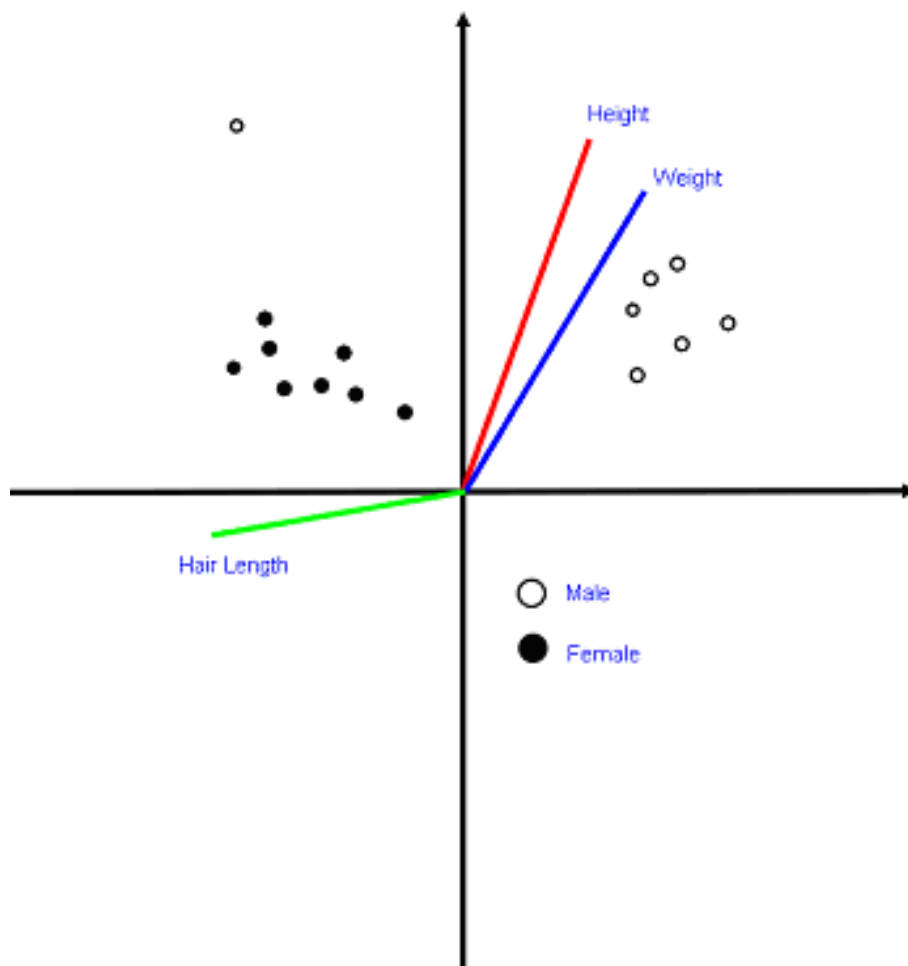


Figure 5.3: An example of a data set to illustrate the properties of a biplot. These data are entirely hypothetical and have been created solely to illustrate these properties. Interpreting this plot, is it possible to see that height and weight are correlated while hair length is not correlated with weight or height. The samples which have been plotted are classified into males and females. It is clear to see that females tend to be shorter and lighter with longer hair. Males show the opposite.

the variance of $\mathbf{Z}_i = \lambda_i$ and the constants $a_{i1}, a_{i2}, \dots, a_{ip}$ are the elements of the corresponding eigenvector that have been scaled such that

$$a_{i1}^2 + a_{i2}^2 + \dots + a_{ip}^2 = 1 \quad (5.2)$$

An important property of the eigenvalues is that they sum to the sum of the diagonal elements of the covariance matrix, since $Z_i = \lambda_i$ and the variance of \mathbf{X}_i is equal to the i th element of the diagonal of the covariance matrix. Thus, the p principal components will describe the variance of the data.

The eigenvectors of \mathbf{X} form the column of an orthogonal matrix \mathbf{V} . These eigenvectors form an alternative basis for the ρ dimensional space within which the samples or fragments are described. Relative to the p dimensional space, the position of these fragments are given by $\mathbf{Z} = \mathbf{X}\mathbf{V}_p$. That is, the best display of the n points is given by the n of the matrix \mathbf{Z} .

The biplot axis represents the variables describing the fragments. These axes are calculates as follows: e_k is a unit vector along the k th coordinate axis in the p dimensional space. The point \mathbf{x} with coordinates (x_1, x_2, \dots, x_p) may be written

$$\mathbf{x} = \sum_{k=1}^p x_k \mathbf{e}_k \quad (5.3)$$

which will be interpolated to

$$\mathbf{x}\mathbf{V}_\rho = \sum_{k=1}^p x_k (\mathbf{e}_k \mathbf{V}_\rho) \quad (5.4)$$

where $\mathbf{e}_k \mathbf{V}_\rho$ is the interpolant of the unit point on the k th axis.

Using both of these features, a biplot is calculated in ρ dimensions.

Search: search1
Database(s): CSD version 5.27 (November 2005)
Restrict Info: No reftype restrictions applied

Filter(s): 3D coordinates determined	R factor <= 0.05
Not disordered	No errors
Not polymeric	No ions
No powder structures	Only Organics


Advanced Options: None

Single query being used. Search will find structures that:

have

Query 1

Query 1



The image shows a search interface for a difluoroalkene fragment. It includes search parameters, filters, and a chemical structure diagram with a 3D model icon.

Figure 5.4: Search information for the fragment difluoroalkene. From this search 33 fragments were extracted.

5.2 Model Examples

5.2.1 Difluoroalkene

Clustering of this fragment is described in [10]. The dendrogram in Figure 5.5 shows the fragments in each cluster. As can be seen in this diagram, the conformation of each of the fragments can easily be distinguished using the fragment viewer. It is hoped that using a biplot, the variables that describe these changes will be easily distinguished. The biplot in Figure 5.5 was created in Matlab where the first two principal components were plotted. As can be seen, it is difficult to distinguish how each of the variables is named. This is a result of a number of problems: 1. there are too many variables (75 variables describing a 6 atom fragment) and 2. these variables are too highly correlated. These two factors produce a plot that is hard to interpret. Nevertheless, it is possible to see that the variables are highly correlated with many variables measuring the same conformational change. When the plot is examined in detail, it is clear that there are variables that are describing specific conformational changes. When the conformations of the fragments in the red, green and blue clusters in Figure 5.5 are examined, it is apparent that the major conformational change is a restriction of the carbon backbone bonded angles. This is reflected in the plot where the axes describing the variables representing the bonded angles are parallel to a trend between these clusters (Figure 5.6). Using the interpolation process, it is possible to assign this conformational change solely to changes in these variables. The majority of the remaining variables are describing the rotational conformational changes. This is where it becomes particularly difficult to uncover which of the variables are representing the conformational change. This is partly the result of the highly redundant data set and partly because of the highly correlated data. It would appear that in order to make biplots a useful tool in *d*SNAP it will be necessary to reduce the number of variables that are required to describe the fragments.

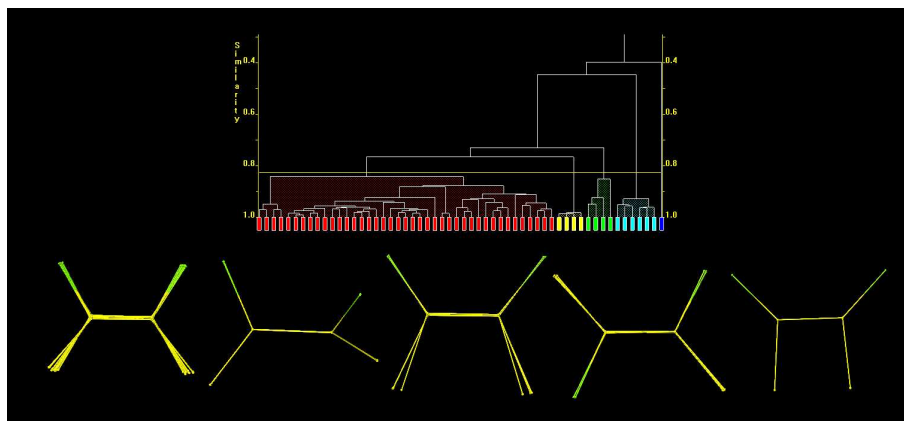


Figure 5.5: Dendrogram of difluoroalkene with the fragments illustrated below. The fragments are arranged such that the content of each cluster is represented from left to right, *i.e.* the left most fragments are from the red cluster.

5.3 Reducing Variables

In this section the number of variables has been reduced by the application of factor analysis. The number of variables had been selected in an identical manner to that described in Chapter 4. It is hoped that the application of factor analysis will remove some of the complication that has resulted in the problems with interpretation of biplots.

5.4 Model examples with reduced variables

5.4.1 3-chlorobut-2-ene-thiolate

These data were first examined in Section 2.2.1. In this section it is demonstrated that there are two major conformational changes within these data. There is a conformational change and a restriction of the bonded angle around the carbon backbone. With reference to Figure 2.3 the conformational changes within these data are clear in most cases. The variables have been treated with factor analysis as described in Section 4.4 and a biplot was drawn from the 46 variables extracted. This biplot is shown in Figure 5.7. The biplot shown in this figure shows that there are two distinct clusters on

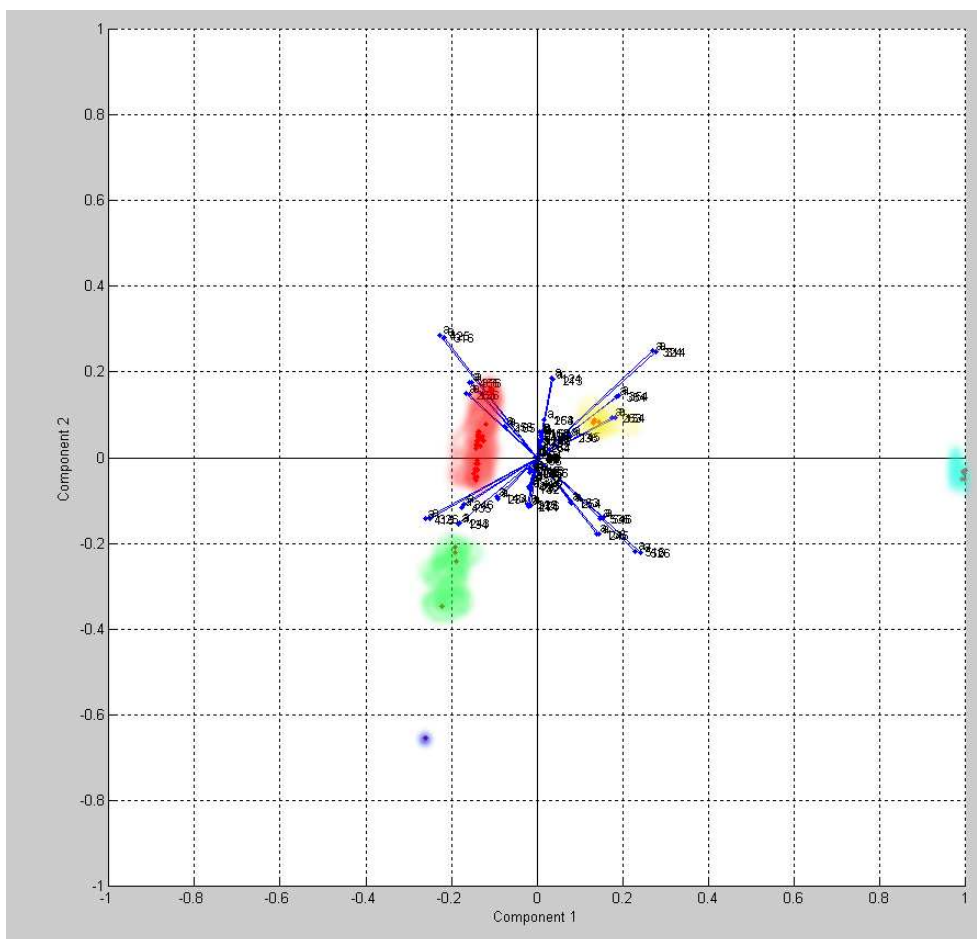


Figure 5.6: Biplot of difluoroalkene with default data matrix where the fragment was defined by total geometries. Colours have been added according to the colours in the dendrogram in Figure 5.5.

either side of the x axis. This change corresponds to the *cis/trans* change in conformation. This is illustrated in Figure 2.3 on page 41. There is also a spread of fragments along the y axis. The fragments with a higher y value have lower value for bonded angle. From the axes representing the variables it is hard to make sense of which of the variables are responsible for the conformational changes. This is not a rare problem with biplots. The analysis has been simplified by the application of factor analysis but there is still the fundamental problem of a large number of highly correlated variables. This plot tells the user as little as the plot in Figure 5.6. It could be possible that the inherent simplicity of these data does not lend itself to this particular type of analysis.

5.4.2 3-aminobutan-2-ol

This fragment was first examined in Section 2.2.2. In summary there are two different types of conformational change in these data. There is a rotational component where the atoms are in different orientation as the result of a rotation around the central bond and there is a constraint on the bonded angle as a result of the chemical context of the fragment. Factor analysis is then applied to these data in a manner described in Section 4.5 and the number of variables have been reduced to 15. With reference to Figure 5.9 it is apparent that there are three different groups of variables.

Approximately following the x axis in Figure 5.9, there are two groups of variables that are highly negatively correlated. The variables that lie along this axis are: a163, a216, a316, a162, a136 and a126 as well as a lesser contribution from a436 and a346. The remaining group of variables are projected orthogonally for the first two groups of variables. This indicates that these variables are uncorrelated with the initial two groups of variables. The variables that lie along this axis are: a415, a365, a256, a265, a356 with a minor contribution from d56 and d46. An illustration of this fragment can be found in Figure 5.8. Uncovering what conformation underlies the pattern in these variables is quite difficult. There should be two different conformational changes as illustrated in Figure 2.10. There does not appear

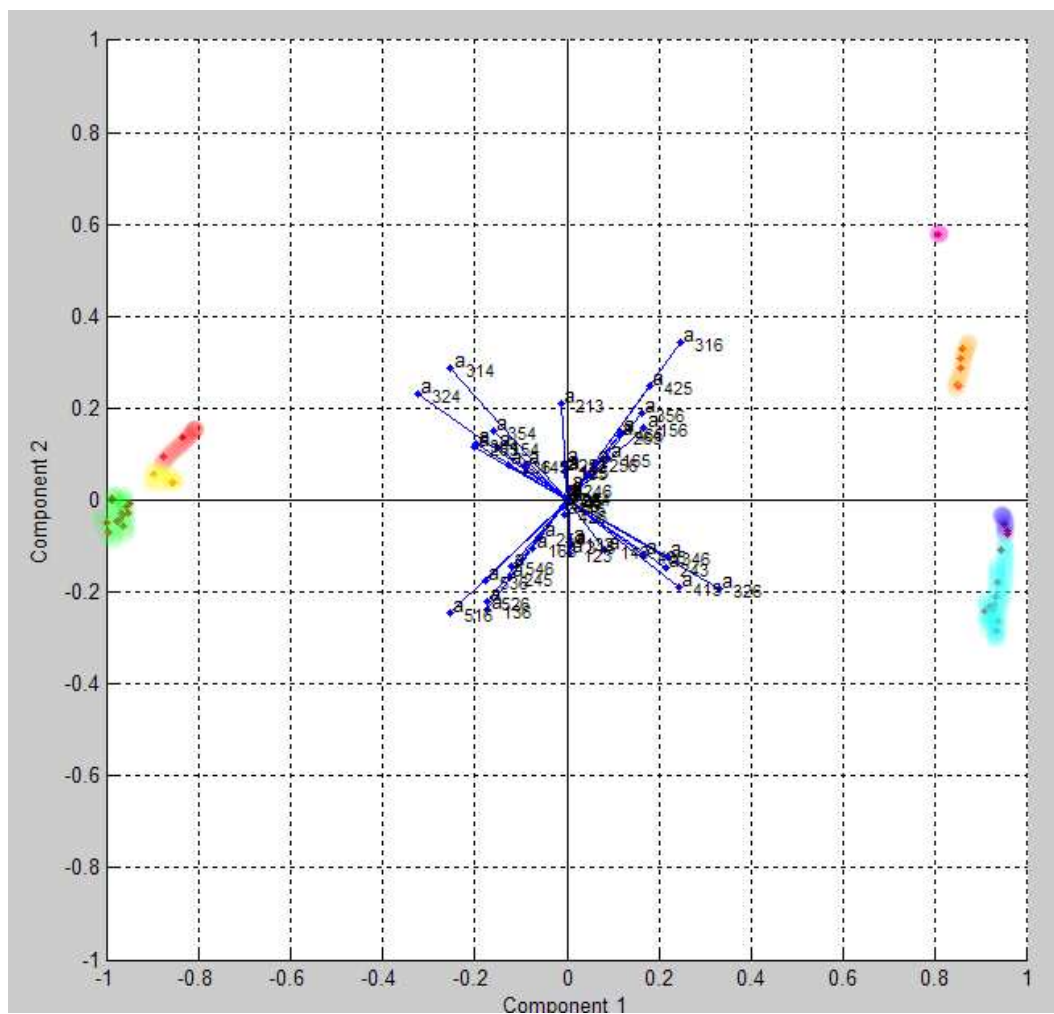


Figure 5.7: Biplot of the fragment 3-chlorobut-2-ene-thiolate where factor analysis was applied to reduce the number of variables required to describe the fragment. The fragments have been coloured such that they match the colours in the dendrogram when the fragment was defined by total geometries.

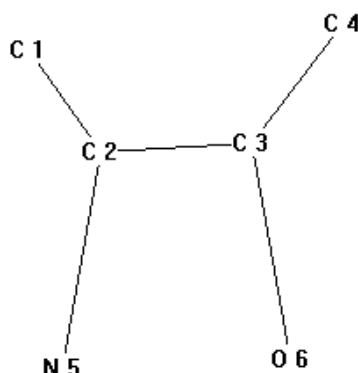


Figure 5.8: Diagram of the fragment 3-aminobutan-2-ol

to a clear distinction between these two different conformational changes based upon the biplot. There is a separation between the fragments that have a very different conformation of carbon atoms. On the right on the plot in Figure 5.9 there are the fragments from the cluster where the carbon atoms are *gauche*, while the fragments to the right of the plot are those where the carbon atoms are in *anti* position. The rotational components are illustrated in Figure 2.11 on page 50 and it is apparent, based upon this biplot, that the rotational component is far more important than the restriction in the carbon backbone of the fragments.

5.4.3 Pentan-2-one

This fragment was described in Section 2.2.3. The major conformational changes within these data are a rotation around the two torsion bonds. The dendrogram coupled with the diagrams in Figure 2.19 shows that five clusters have formed when the dendrogram was cut at this level. When factor analysis was applied to these data 31 variables were extracted with a factor loading of greater than $|0.95|$. When these variables were analysed in *dSNAP*, the analysis yielded clusters with good agreement with the clusters originally calculated when the fragments were defined by total geometries. These variables were then displayed as a biplot. This biplot is shown in Figure 5.10. As can be seen in this figure, the fragments form clear clusters and there are a number of variables with varying degrees of correlation. In terms of the

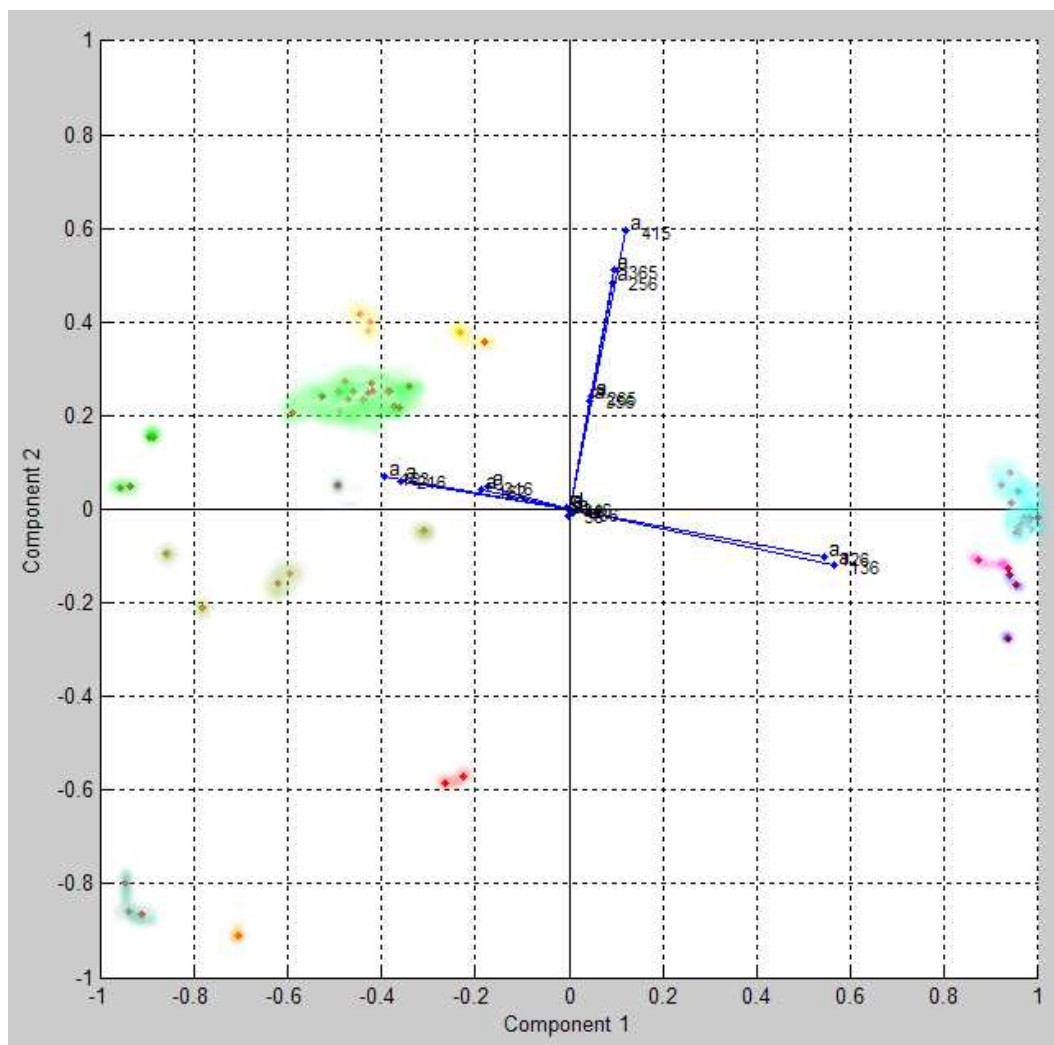


Figure 5.9: Biplot of the fragment 3-aminobutan-2-ol where the number of variables has been reduced by the application of factor analysis. The fragments have been coloured to match the clusters that were found in when the fragment was defined by total geometries.

fragments plotted onto the biplot, there are three clear clusters formed. The colouring of the clusters correspond to the colours displayed in the dendrogram 2.19. The fragments in the biplot therefore represent the conformation displayed in the Newman projections in Figure 2.19. The two largest clusters are the red and yellow clusters. The major difference is a difference in the conformation of the second torsion bond of the fragment. As a result of the initial definition of the geometry of the fragments, a large number of the variables within these reduced data are measuring this change. The variables which have the largest contribution, with reference to the biplot in Figure 5.10, are a456 and a546. Both of these angles are measuring a change across both of the torsion angles that are present in this fragment. This make the plot difficult to understand but it appears that most of the variables that are plotted along the x axis represent variables that are representing a change in the second torsion angle. The variables which are approximately along the y axis are representing variables that are measuring a change in the first torsion angle.

5.5 Conclusions

To say that there is clear assignment of features in the biplot to conformational changes for the fragments is unwise. As discussed in Chapter 2 each variable does not distinctly describe a single change in conformation. The nature of these data results in biplots that are inherently difficult to analyse. All of the biplots above show some structure and it is certainly possible to attribute this structure to the underlying conformational changes in these data. The difficulty arises when the biplot is used to aid the interpretation of the clustering. Using the biplot alone, is it difficult to assign structural meaning to the fragments that make up these data. It is possible to understand how variables are affecting the clustering but this is only possible once the clusters are understood from a structural point of view.

Nevertheless, the application of biplots to the data produces interesting results. Certainly the biplot is much easier to understand than the correlation matrix and gives a user of *dSNAP* a useful tool to understand what variables

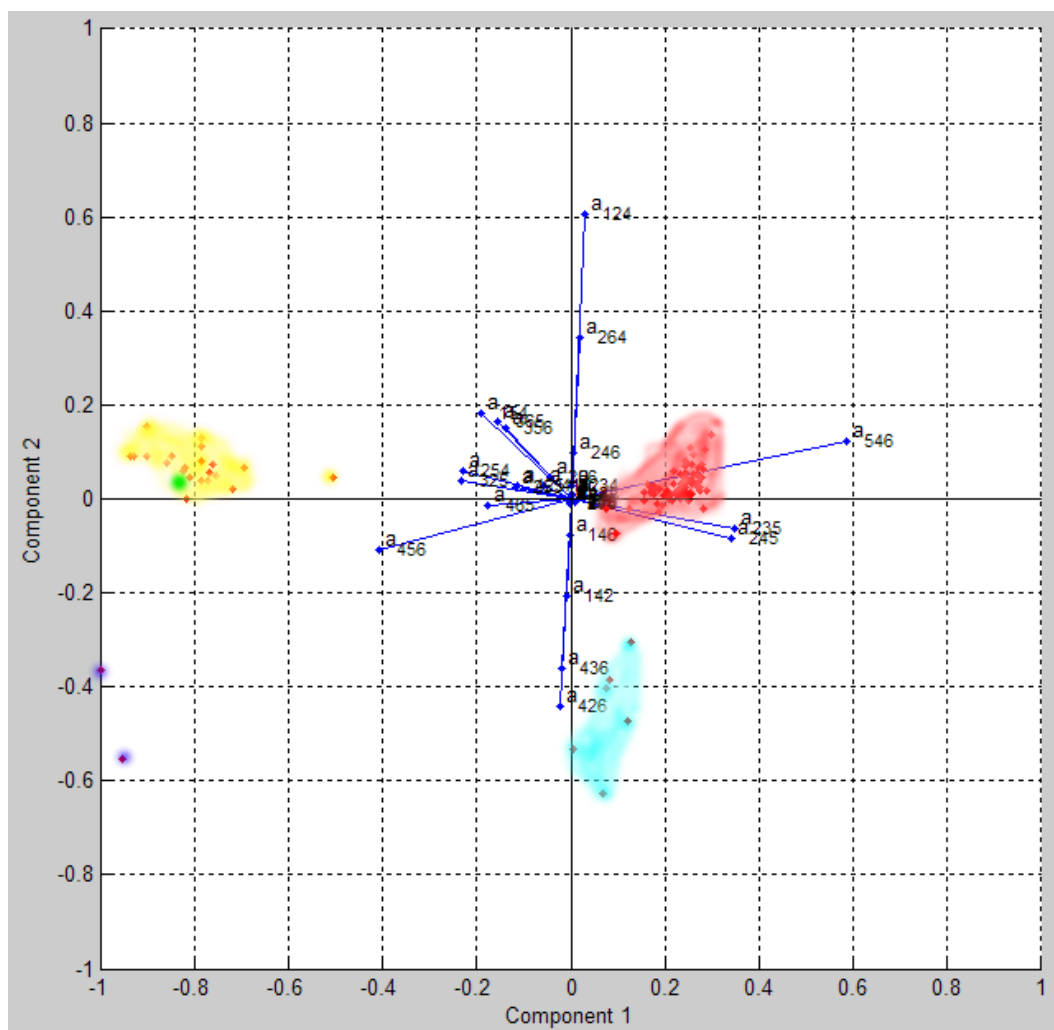


Figure 5.10: Biplot of the fragment pentan-2-one where factor analysis has been used to reduce the number of variables before analysis. The colours that have been applied to the fragments such that they match the colours of the clusters that the fragments are in when the clusters was defined by total geometries.

are affecting the formation of clusters even though as a tool for understanding the structural reason for the formation of cluster it is of limited use.

Chapter 6

Sparse principal components analysis

DSPCA[21] or sparse principal components analysis was investigated to see if this method will aid in the identification of important variables that give rise to the formation of clusters. Principal components analysis [36] is an orthogonal linear transform that transforms a set of data to a new coordinate system such that the greatest variance of the data in projection lies on the first coordinate (the first principal component), the second greatest variance on the second coordinate etc. For a data matrix \mathbf{M} , with zero empirical mean (i.e. the mean of the distribution has been subtracted from the data set), the PCA transformation is given by:

$$\mathbf{Y}^t = \mathbf{M}^t \mathbf{W} = \mathbf{V} \mathbf{\Sigma} \quad (6.1)$$

where $\mathbf{V} \mathbf{\Sigma} \mathbf{M}^t$ is the SVD of \mathbf{M}^t .

The procedure involves calculating the eigenvalues and associated eigenvectors

Given a set of points in Euclidean space, the first principal component is the eigenvector with the largest eigenvalue. This corresponds to a line that passes through the mean and minimizes the sum squared error with those points. The second principal component corresponds to the same concept after all correlation with the first principal component has been subtracted

out from the points. Each eigenvalue indicates the portion of the variance that is correlated with each eigenvector. Thus, the sum of all the eigenvalues is equal to the sum squared distance of the points with their mean divided by the number of dimensions. PCA essentially rotates the set of points around their mean in order to align with the first few principal components. This moves as much of the variance as possible into the first few dimensions. The values in the remaining dimensions, therefore, tend to be highly correlated and may be dropped with minimal loss of information. PCA is often used in this manner for reducing data dimensionality. It is the optimal linear transform for keeping the subspace that has largest variance.

The disadvantage of PCA is that the principal components are usually a linear combination of all the variables i.e. all the weights in the linear combinations (called loadings) are non zero. In many applications, however, the coordinates axes have a physical interpretation e.g. they could be a specific geometric parameter. In these cases the interpretation of the principal components could be facilitated if these components involved very few non-zero loadings. In sparse principal components analysis [21] we sacrifice some of the explained variance and orthogonality in exchange for a situation in which most of the weights are either zero or very small.

Let $\mathbf{A}_{(n \times n)}$ be a symmetric covariance matrix from which we want the sparse principal components. Let k be an integer such that $1 \leq k \leq n$. We want to maximize the variance of a vector \mathbf{x} while constraining its cardinality:

$$\begin{aligned} &\text{Maximize} && \mathbf{x}^T \mathbf{A} \mathbf{x} \\ &\text{Subject to} && \|x\|_2 \\ &\text{and} && \text{Card}(x) \leq k \end{aligned}$$

This is a NP Hard problem which means that there is no easy solution in a finite time. This problem gets relaxed by by d'Apremont *et al.*[20, 21] and it is now formulated as follows:

Given a matrix \mathbf{A} we wish to decompose it into factors with a target sparsity k . To do this we:

$$\begin{aligned}
&\text{Maximise} && \text{Tr}(\mathbf{A}\mathbf{X}) \\
&\text{Subject to} && \text{Tr}(\mathbf{X}) = 1 \\
&&& \mathbf{1}^t |\mathbf{X}| \leq k \\
&&& \mathbf{X} \succeq 0
\end{aligned}$$

where

$$\mathbf{X} = \mathbf{x}\mathbf{x}^T$$

A matrix $\mathbf{X}_{(n \times n)}$ is positive definite if for any non-zero vector $\mathbf{z}(n)$ with real entries (We write this $z \in R^n$)

$$\mathbf{z}^t \mathbf{M} \mathbf{z} > 0$$

If

$$\mathbf{M} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

and if we take any vector

$$\mathbf{z} = \begin{pmatrix} z_1 \\ z_2 \end{pmatrix}$$

then

$$\mathbf{z}^t \mathbf{M} \mathbf{z} = (z_1 z_2) \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = z_1^2 + z_2^2$$

Since the vector \mathbf{z} is non-zero, either $z_1 > 0$ or $z_2 > 0$ so $z_1^2 + z_2^2 > 0$ so $\mathbf{z}^t \mathbf{M} \mathbf{z} > 0$

A matrix is Hermitian if it is a square matrix with complex entries which is equal to its own conjugate transpose *i.e.* the element in the i th row and j th column is equal to the complex conjugate (denoted with a $*$) of the element in the j th row and i th column, for all indices i and j :

$$a_{ij} = a_j^* i$$

e.g. the matrix

$$\begin{pmatrix} 3 & 2+i \\ 2-i & 1 \end{pmatrix}$$

is Hermitian.

A Hermitian matrix \mathbf{M} is positive semidefinite if

$$\mathbf{z}^t \mathbf{M} \mathbf{z} \geq 0$$

For *any* matrix \mathbf{M} , the matrix $\mathbf{M}^* \mathbf{M}$ is positive semidefinite

Notation: The constraint that a matrix \mathbf{M} is positive semidefinite is written

$$\mathbf{X} \succeq$$

The trace of a matrix is the sum of its diagonals and is written $\mathbf{Tr}(\mathbf{M})$ *e.g.*

$$\mathbf{Tr} \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} = 5$$

The cardinality of the vector is the number of non-zero components. It is written $\mathbf{Card}(\mathbf{z})$. *e.g.* for the vector

$$\mathbf{Card}(\mathbf{z}) = \begin{pmatrix} 0.5 \\ 0 \\ 1.5 \end{pmatrix} = 2$$

For a matrix rather than a vector the cardinality is the number of nonzero

coefficients.

The column rank of a matrix \mathbf{M} is the maximal number of linearly independent columns of \mathbf{M} ; the row rank is the maximal number of linearly independent rows of \mathbf{M} . Since the column rank and the row rank are always equal, they are simply called the rank of \mathbf{M} . We write this as $Rank(\mathbf{M})$.

A vector of 1s is written $\mathbf{1}$

6.1 Method

Analysis was carried using the implementation of DSPCA created by A. d’Aspremont *et al* [21, 20]. The following settings were applied:

Input	Value
algo	1 (full eigenvalue decomposition)
gapchange	0.05
rho	0.5
info	1
maxiter	1000

Table 6.1: Settings for sparse PCA. algo controls the method for computing the matrix exponential. Gapchange is the target reduction in duality gap. Maxiter is the maximum number of iterations and ρ is a parameter controlling sparsity. Info controls the verbosity of the reporting.

The program DSPCA was run in MATLAB using following the user guide provided from [20]. The dominant eigenvector from this analysis was used to select important variables for each of the examples and are tabulated in Appendix D.

6.2 Model examples

6.2.1 Difluoroalkene

This fragment bears a close resemblance to the fragment 3-chlorobut-2-ene-thiolate described in Section 2.2.1. The clustering is also described in [10].

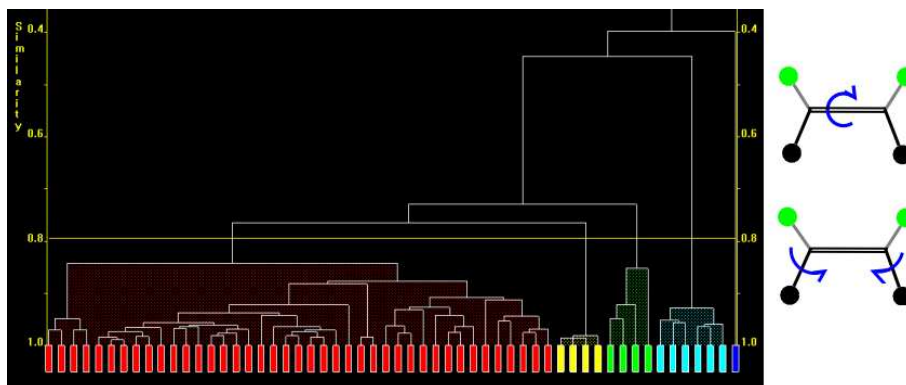


Figure 6.1: Dendrogram of difluoroalkene with diagram of the fragment. For this dendrogram the fragment was defined by total geometries.

Within these data there are two major conformational changes: A *cis/trans* conformational change and a restriction of the bonded angles dictated by the molecular context from where the fragment was derived from. The red, yellow, green and blue clusters are in *trans* conformation. The difference between these fragments is the chemical context that the fragment finds itself in. The red cluster is in either the backbone of a molecule or in a greater than six member ring. The yellow cluster is in an eight member ring but the twisted double bond conformation is the result of a mis-classification of the single and double bonds when the structure was solved. The green and blue fragments have their backbones constrained in either a five member ring or a four member ring. The remaining cyan cluster is in *cis* conformation and the backbone is in the backbone of the original fragment.

The fragment had its geometry described using total geometries. Sparse principal components analysis was applied to these variables with the intention of identifying the key variables that are causing the formation of these clusters. By applying sparse principal components analysis the number of variables has been reduced from 75 describing the fragment with total geometries to eight variables. These variables were chosen from the first eigenvector calculated during sparse principal components analysis. This can be seen in Table D.1 on page 162. The variables deemed important were selected by thresholding those variables that had an absolute value of greater than 0.2.

These variables were then extracted from the original data matrix, tabulated and used as the input for *d*SNAP. The variables extracted are a314, a316, a415, a516, a324, a326, a425, a526. In Figures 6.2 and 6.3 it is clear that when the number of variables has been reduced by the application of sparse principal components analysis, the clusters generated in *d*SNAP using these variables matches the clusters generated when the fragments were described by total geometries. This shows that applying sparse principal components analysis to these data has selected variables that accurately describe the geometries of the fragments.

6.2.2 3-aminobutan-2-ol

This fragment was described in Section 2.2.2. There are essentially 2 major conformational changes; there is a rotational component involving a rotation around the central bond along with the relative chirality of the fragments and there is also a restriction in the bonded angles owing to the molecular context from where the fragments were derived from. These conformational changes are illustrated in Figure 2.10. Figure 6.4 shows the different conformations of the fragments within these data.

When sparse principal components analysis was applied to the 75 variables that describe the fragment in total geometries, 11 variables were deemed to be significant and were extracted. That is, the absolute value for the first eigenvector was greater than 0.2. This can be seen on Table D.2 on page 164. These variables were then tabulated and was used as an input for *d*SNAP. The variables selected were a415, a126, a425, a134, a136, a435, a143, a146, a154, a164. The output from *d*SNAP was then analysed to compare the clustering from the original analysis described in Section 2.2.2 with the output from sparse principal components analysis. As can be seen in Figure 6.6, the dendrograms are quite different in general appearance. However, it can be seen from Figure 6.5, the clusters of fragments derived from total geometries are partially preserved even when the number of variables has been reduced by applying sparse principal components analysis. When analysis was carried out on the fragments described using key variables identified using DSPCA,

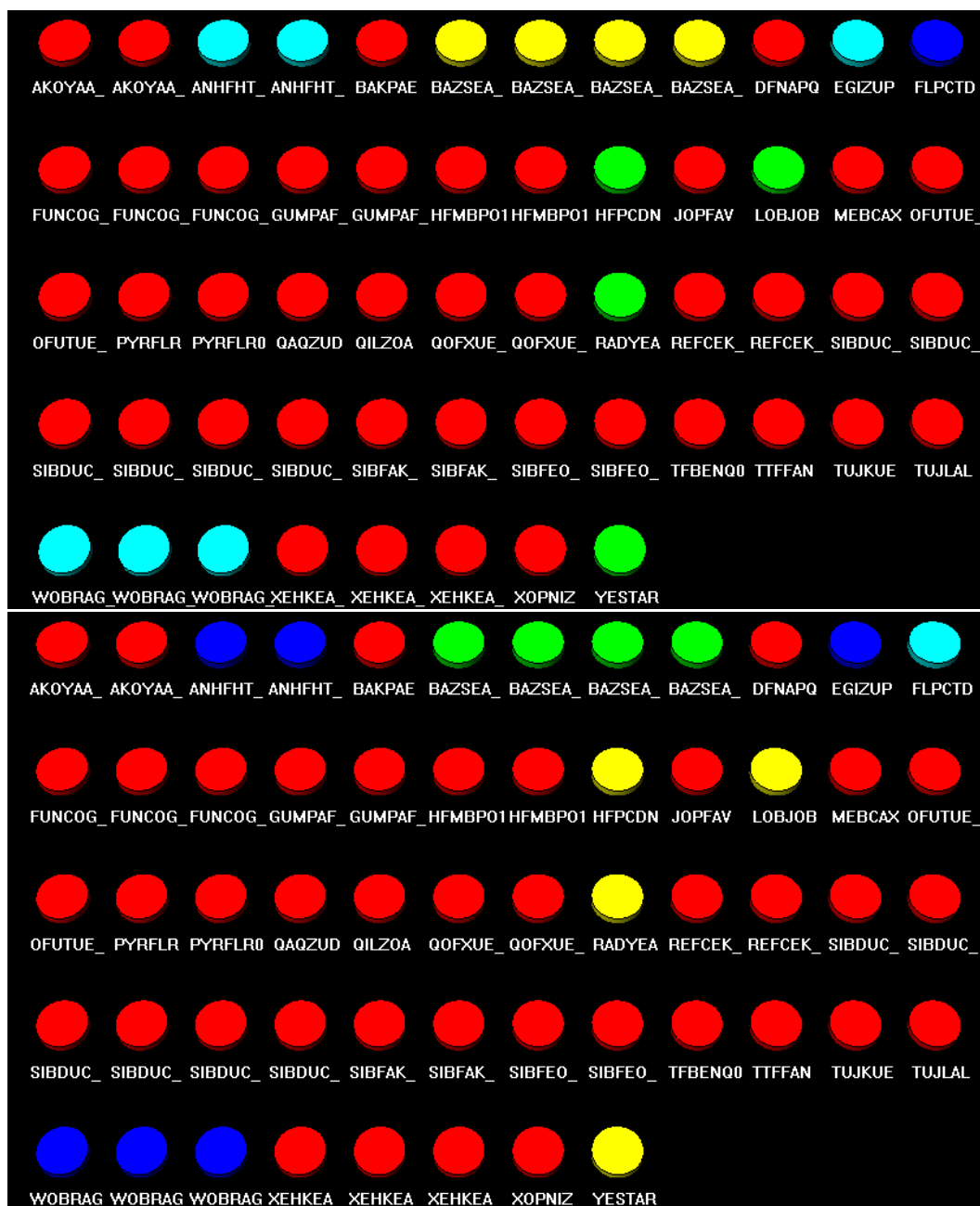


Figure 6.2: Cell displays of the fragment difluoroalkene with the fragment described by total geometries on the top and with the variables reduced by the application of sparse principal components analysis on the bottom.

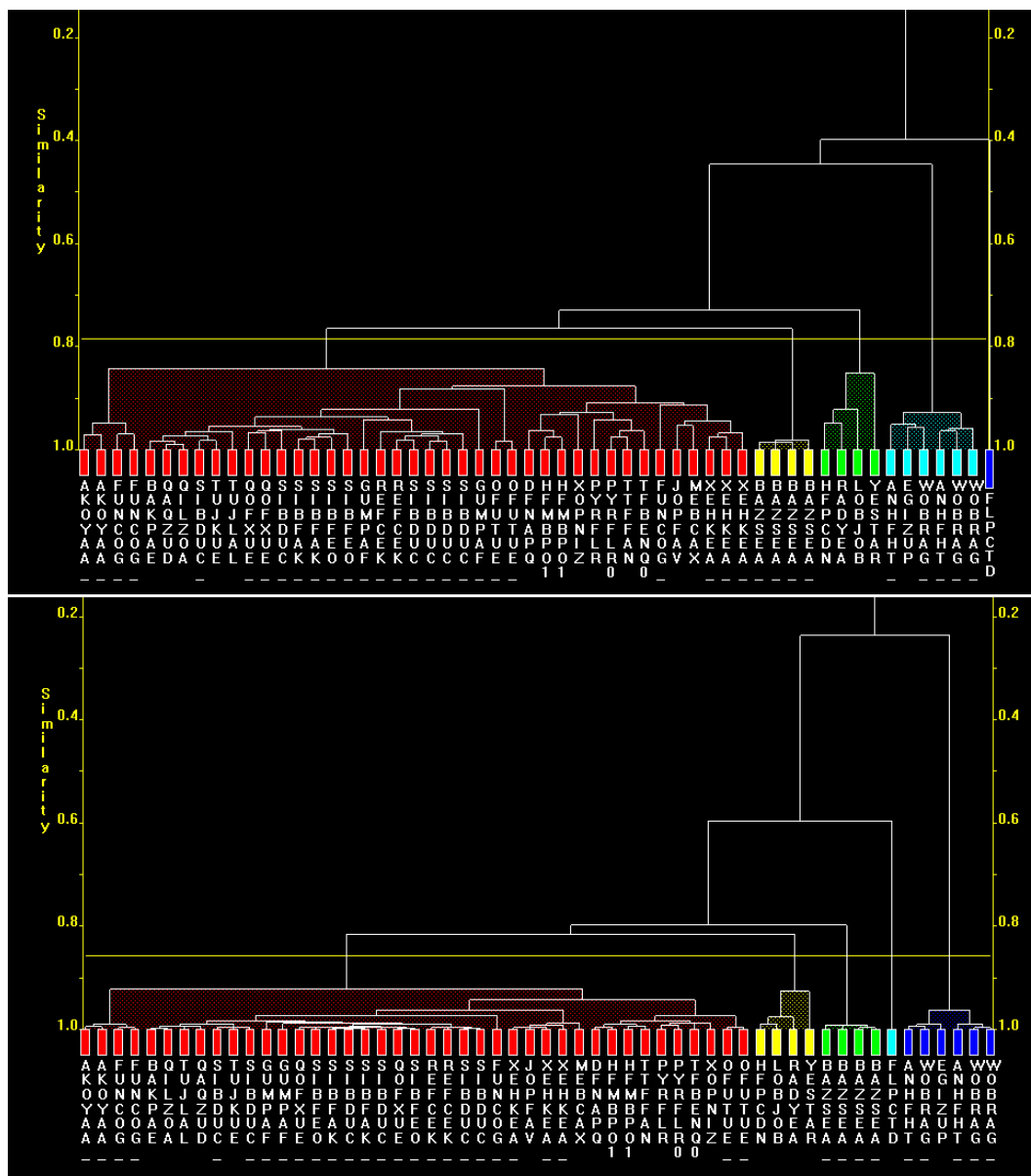


Figure 6.3: Dendrograms of the fragment difluoroalkene with the fragment described by total geometries on the top and with the variables reduced by the application of sparse principal components analysis on the bottom.

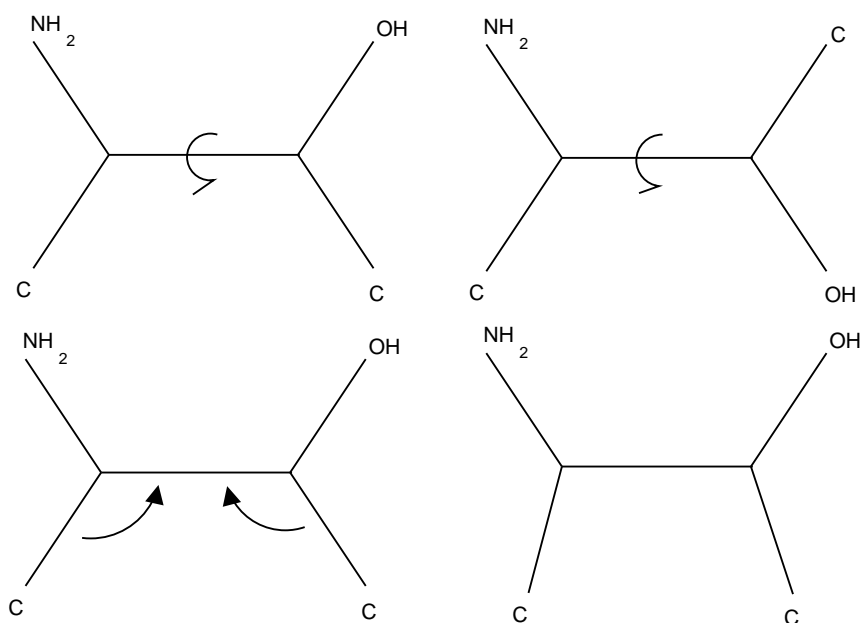


Figure 6.4: Expected geometric changes in the fragment 3-aminobutan-2-ol. Also there are 2 chiral centres which should also be found in S^* and R^* conformation.

it was found that cluster formation was fairly rational with respect to the broad conformations of the fragments. It was found that there was a loss of fine detail that separates some of the minor conformational changes. This minor loss of information could account for the different clustering. A conclusion to this initial exercise is that in order to maintain the precision of the clustering it may be necessary to analyse more eigenvectors from the output of sparse principal components analysis.

6.2.3 Pentan-2-one

The geometry of this fragment was described in Section 2.2.3 where the fragment was described using total geometries. A summary of the conformations can be found in Figure 2.19. In summary the major conformational changes within these data are a rotation around both of the backbone torsion angles as illustrated in Figure 2.19.

Sparse principal components analysis was applied to the 75 variables de-

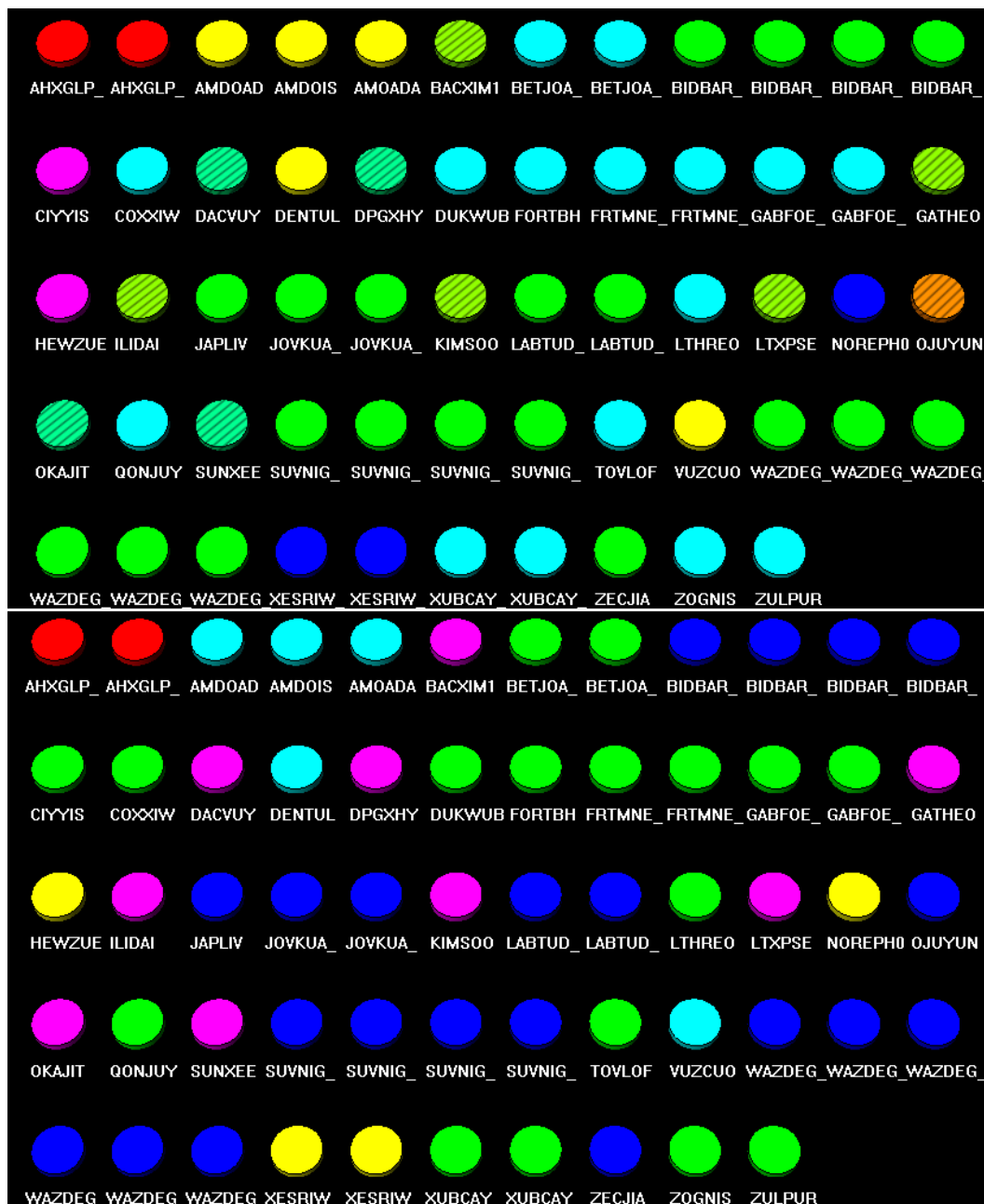


Figure 6.5: Cell display of 3-aminobutan-2-ol with the fragment defined by total geometries on top and the cell display with the variables describing the fragment reduced by the application of sparse PCA on the bottom.

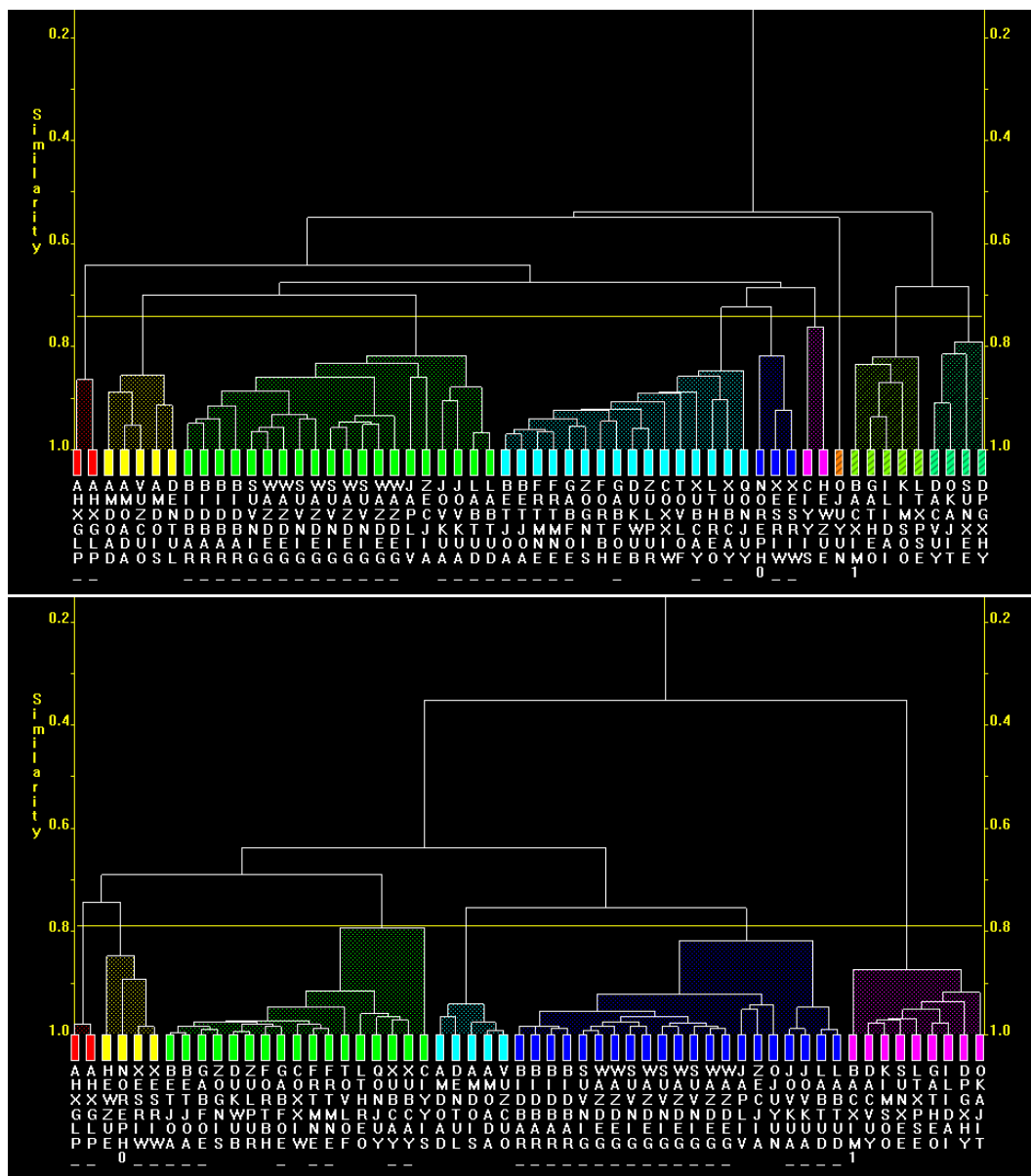


Figure 6.6: Dendrogram of comparing the fragment 3-aminobutan-2-ol when the fragment was described by total geometries on the top and with the variables reduced by the application of sparse PCA on the bottom.

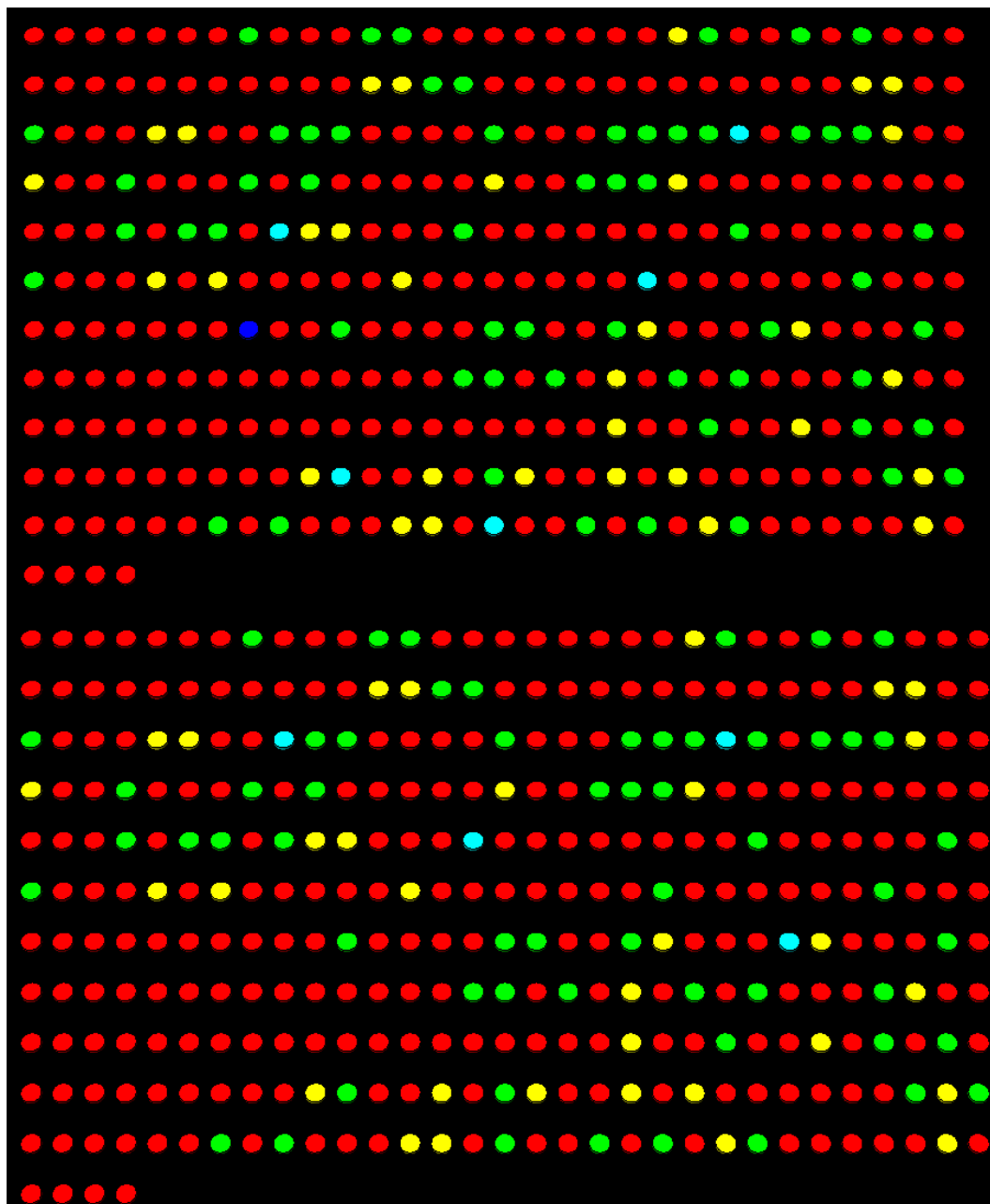


Figure 6.7: Cell display comparing the fragment pentan-2-one where the fragment was defined by total geometries (top) and when the number of variables are reduced by the application of sparse principal components analysis (bottom)

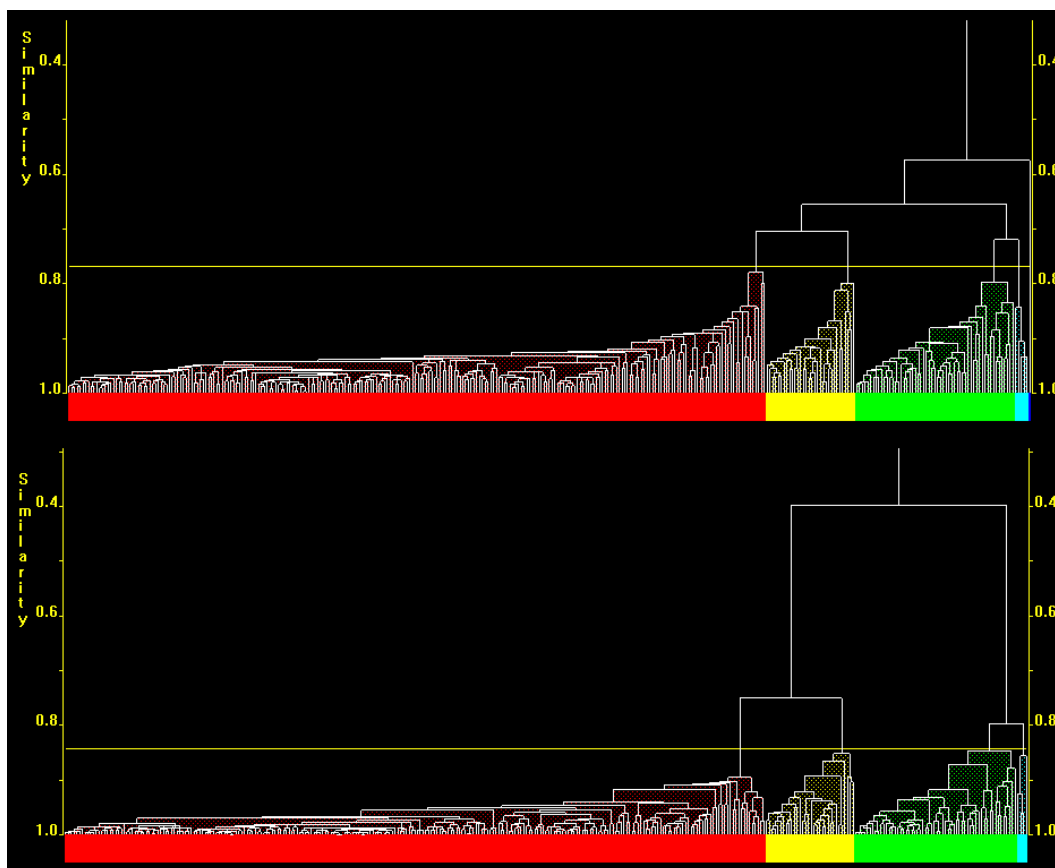


Figure 6.8: Dendrogram comparing the fragment pentan-2-one where the fragment was defined by total geometries (top) and when the number of variables are reduced by the application of sparse principal components analysis (bottom)

scribing the total geometries of the fragment pentan-2-one and 7 variables were extracted. An important variable was deemed to have an absolute value for the first eigenvector greater than 0.2. The initial eigenvector can be seen in Table D.3 on page 166. These variables were then tabulated and used as the input for *d*SNAP. The variables extracted are a135, a235, a536, a145, a245, a546, a456. When this output of *d*SNAP was compared with the original output when the fragment was defined with total geometries there is good agreement between the 2 cell displays in Figure 6.7. The dendrograms in Figure 6.8 also show that there is good agreement between the 2 different definitions of the fragment. This shows that the 7 variables extracted from by the application of sparse principal components analysis have accurately clustered the fragments into groups with similar conformation.

6.3 Conclusions

This section looks into the application of sparse principal components analysis with the aim of reducing the number of variables required to describe the conformation of the fragments. The 3 examples above indicate that the application of sparse principal components analysis to the variables describing the fragments in question has significantly reduced the number of variables. In the case of pentan-2-one the number of variable has been reduced by an order of magnitude. This indicates that sparse principal components analysis in the form of DSPCA [21] has successfully reduced the number of variables required to describe the geometry of the fragment in these examples. While there is not exact agreement between the output from *d*SNAP when the fragments were defined with total geometries and variables extracted by applying sparse principal components analysis there is enough to show that this method is extremely promising. It would be of interest to see if DSPCA could be extended to extract more than a single eigenvector. If this were to be true then sparse principal components analysis may well be of use as an aid for understanding the formations of clusters in *d*SNAP.

Chapter 7

Conclusions

This thesis has concentrated on understanding the nature of the geometric definition of substructural chemical fragments and their application to the program *dSNAP* and with investigating methods of reducing the number of variables required to describe the conformation of fragments.

Broadly speaking the program groups the fragments which have been mined from the Cambridge structural database into groups that have a similar conformation or shape. The definition that describes this conformation is vital to the accurate function of the program. In Chapter 2 there is an examination into various geometric definitions that can be applied to fragments. Each of these definitions has their merits but they can be broadly grouped into bonded and non-bonded definitions. The bonded definitions are those that only involve variables that directly measure the bond distances, the bond angles and torsions angles. With reference to the model examples in Chapter 2: When the fragments were defined by bonded variables the fragments were not grouped into clusters made up of fragments with distinct conformations. There were a few interesting characteristics that were exposed when clustering using bonded definitions. One such example is the fragment pentan-2-one where there were some fragments that had shorter bond lengths that were masked when using total geometries. When the fragments were defined using non-bonded variables most of the fragments formed clusters that were made up of fragments with similar conformations. This indicates

that these definitions have done an adequate job of classifying the fragments into sensible clusters containing fragments with similar conformations that are discrete in conformation from the other clusters. When the different non-bonded definitions are examined there is not much difference between them. The non-bonded definitions are total geometries, all angles and all distances. In theory either of these definitions could be used in *dSNAP* to define the conformation of the fragments. The advantage of total geometries is that both the distances and angles can be examined, particularly in variable space where individual variables can be compared with each other.

One of the major drawbacks to this definition is that the variables are highly redundant. For example to describe the position of three points in space it requires six variables. This proves to be a problem when attempting to detect specific changes in conformation by observing changes in the variables with these data. Triangles (Chapter 3) discuss the possibility of summarising a number of variables by describing collections of variables as triangles. In the example above the three points could be described using the area of a triangle. Initially simulations of the area of triangles were carried out. These show that the area of a triangle will change with different rotations around a torsion angle. When a single torsion angle was simulated it was relatively easy to see the relationship between the torsion angle and triangle area. When two torsion angles were simulated the relationship between torsion angle and triangle area is much more complicated. With reference to Figure 3.6 it is apparent that for a given area of triangle there are many possible torsion angles that could give rise to that area. This could well be one of the reasons why when the fragment pentan-2-one was defined by triangles in Section 3.2.5 there was little agreement between the output from *dSNAP* when the fragment was defined with this definition and when the fragment was defined with total geometries. There are also issues with the suitability of triangles to describe the geometries of certain fragments. Also, in this implementation the triangles have to be calculated manually which is against the principal where the geometry of the fragments should be calculated automatically.

In an attempt to reduce the number of variables required to accurately

describe the geometry of the fragments factor analysis was applied. Factor analysis attempts to extract latent variables that describe hidden features of the dataset. These hidden variables could represent different changes in conformation but it is not clear which factor is measuring which change in conformation. Nevertheless by selecting variables that are strongly related to these factors it was possible in some cases to reduce the number of variables required to describe the geometry of the fragment. However there are some situations where a relatively low percentage of the variance within a given dataset can be explained by applying factor analysis. There is also an issue where the variables are all interrelated. This could be one of the reasons why it is difficult to assign a name to a given factor. It was hoped that each factor could be named after a different conformational change. While this is unfortunate, the application of factor analysis had significantly reduced the number of variables required to describe the fragments.

Biplots are a method to display both the variables and the samples, or in this case fragments, to be displayed in a single plot. With some basic explanation it is easy to interpret the plot. One of the problems that became apparent is that as the number of variables increases the biplot becomes increasingly difficult to understand. This is a particular problem in this context as the number of atoms increases the number of variables increases of the order n^3 where n is the number of atoms. To this end factor analysis was applied to reduce the number of variables required to describe the conformation of the fragments. With these reduced data the biplots became much easier to understand. Using a biplot it is possible in effect to view the entire correlation matrix combined with an indication of the variability of each variable. Utilising the properties of biplots it is possible to select groups of variable which are differentiating between conformations of fragments. This ultimately is the reason why biplots were utilised in this research. By examining the biplots in Chapter 5 it is possible to see the groups of variables that are differentiating between the different conformations of fragments.

Sparse principal components analysis was applied to the input data for *d*SNAP where the fragments were described by total geometries. Variables that were deemed to be significant were extracted by thresholding. That

is, variables that were above a certain threshold value were extracted and used as an input for *d*SNAP. By comparing the output from the fragments described by total geometries and when variables were extracted using sparse principal components analysis there is good agreement between the 2 outputs. It should also be noted that the number of variables extracted using this method is much lower when compared with the other methods of reducing the number of variables above. If the implementation of the method could be extended to produce more than 1 eigenvector then this could lead to a useful method for identifying the key variables describing the formation of clusters in *d*SNAP

7.0.1 Future Work

The implementation of multivariate statistics to *d*SNAP for the purpose of systematically reducing the variables would be a useful aid for a user of *d*SNAP. If the fragment is described using total geometries, the number of variables required to describe a medium to large fragment are vast and highly redundant. When examining thousands of variables with the aim of understanding why the clusters of fragments have formed, statistical methods that ease this burden are essential if the program is to remain user friendly. The implementation of factor analysis and sparse principal components analysis in *d*SNAP would be a huge advantage. However, it would not be prudent to apply factor analysis and sparse principal components analysis to total geometries before performing the standard analysis that *d*SNAP carries out. As shown in the chapters above, there are occasions where the datasets reduced by applying these methods have not faithfully reproduced the clusters generated when the fragments were defined using total geometries. Nevertheless, as post cluster analysis processing, the above methods could drastically reduce the time taken to understand the clustering.

These statistical methods would become much more accessible if they were combined with biplots. Biplots are relatively easy to understand and give a much more informative overview of a correlation matrix than would be possible by viewing a large correlation matrix. If multivariate statistical

analysis was employed before constructing the biplot, it may be possible to reduce a large, highly correlated dataset to a few key variables. These variables might not necessarily be the sole variables that are causing clusters of fragments to form but could be regarded as the most representative variable of a number of highly correlated variables. For example, if after factor analysis a variable with the highest loading was chosen to represent the factor, it could be possible to generate a clear and easy to interpret biplot that could significantly ease the burden of interpretation on the user.

It may also be possible in the case of very large fragments to use the definition of all distances opposed to total geometries. This would reduce the number of variables describing the fragment and as demonstrated in Chapter 2 there is good agreement between this definition and total geometries. Unfortunately, the use of the area of triangles to describe the shape of fragments is too subjective and does not perform adequately well to be implemented in *d*SNAP.

Appendix A

Geometric analysis

Table A.1: Table that describes whither the fragment pentan-2-one has fallen into the same cluster as when the fragment was described by total geometries.

refcode	TG	Angles	Distances	Bonded	Torsion
AVAGIN	A	A	A	A	A
YAXGAG	A	A	A	A	A
BEWHUG	A	A	A	A	A
JATXIL	A	A	A	A	A
JUJJUT	A	A	A	A	A
YEQBUR	A	A	A	A	A
CITDOY	A	A	A	A	A
HABZUF	A	A	A	A	A
JEYJED_02	A	A	A	A	A
TIVKUE	A	A	A	A	A
GIZRIQ	A	A	A	A	C
PAPDUG	A	A	A	A	C
NACJIK	A	A	A	A	A
HERBAS	A	A	A	A	A
HEDMOT	A	A	A	A	A
RIBVUU_01	A	A	A	A	A
DEZVOT	A	A	A	A	A
DICREM	A	A	A	A	C
EABZEM	A	A	A	A	C
XOCXUL_02	A	A	A	A	C
XOCXUL_01	A	A	A	A	A
JEYJED_01	A	A	A	A	C
XOCXUI01_01	A	A	A	A	C
XOCXUI03_01	A	A	A	A	C
XOCXUI01_02	A	A	A	A	A
XOCXUI03_02	A	A	A	A	A
XIDNOO	A	A	A	A	C

Continued on next page

Table A.1 – continued from previous page

refcode	TG	Angles	Distances	Bonded	Torsion
XOCXUI02	A	A	A	A	A
PILHAU	A	A	A	A	A
QELMOJ	A	A	A	A	A
GEKDAC	A	A	A	A	A
JEKPIY	A	A	A	A	A
RIBVUU_02	A	A	A	A	A
ROLNIP	A	A	A	A	C
UBUNIO	A	A	A	A	A
EFAXAK	A	A	A	A	C
EFAXEO	A	A	A	A	A
VAPQEJ	A	A	A	A	A
ZIXTAB	A	A	A	A	A
JAKGEH	A	A	A	A	A
JEYJED_03	A	A	A	A	A
BEZWEI	A	A	A	A	C
GALCAX_01	A	A	A	A	A
GALCAX_02	A	A	A	A	A
GALCAX02	A	A	A	A	A
ZIKTUI	A	A	A	A	C
FOFKUH	A	A	A	A	A
REFREZ	A	A	A	A	A
GALCAX01	A	A	A	A	A
LIDJOY_01	A	A	A	A	A
LIDJOY_03	A	A	A	A	A
HOTSIS	A	A	A	A	A
PIWJIO	A	A	A	A	A
FAKZEX	A	A	A	A	C
YASVEU	A	A	A	A	C
KUFNII	A	A	A	A	C
LELGUE	A	A	A	A	C
LUCGAR	A	A	A	A	A
NEMPUR_01	A	A	A	A	A
NEMPUR_02	A	A	A	A	A
EZOMEL	A	A	A	A	C
LUNNIR_01	A	A	A	A	A
LUNNIR_02	A	A	A	A	A
SOHTEO	A	A	A	A	C
HEYMAA	A	A	A	A	A
JIDHUZ	A	A	A	A	A
WINNEM	A	A	A	A	C
QAHJUF	A	A	A	A	A
YIDMED	A	A	A	A	C
BANVAO	A	A	A	A	C
SOCJEZ	A	A	A	E	C
VUCSOB	A	A	A	E	A
LELDIP	A	A	A	E	A
TOZHOF	A	A	A	F	F

Continued on next page











Table A.1 – continued from previous page

refcode	TG	Angles	Distances	Bonded	Torsion
YOLMUH	A	A	A	A	C
CEWWAC01	A	A	A	E	A
PEZNAK_01	A	A	A	E	A
PEZNAK_02	A	A	A	E	A
ECETUB	A	A	A	A	C
TEJSEH	A	A	A	E	A
WEFMEA	A	A	A	E	C
RUVSIK	A	A	A	A	A
ZUWPAI	A	A	A	B	A
BULCEQ	B	D	B	A	B
TEKWEL	B	D	B	A	B
RAKFEP	B	D	B	A	C
RIZWUS	B	D	B	A	C
FEXFEU_02	B	D	B	A	C
LIDJOY_02	B	D	B	A	C
LIDJOY_04	B	D	B	A	C
CEGDOH	B	D	B	A	C
YEHRIM	B	D	B	A	C
KANDUY	B	D	B	A	C
HALDOL01	B	D	B	A	B
LOZDIN	B	D	B	A	B
MAJUSB	B	D	B	A	C
WOMREV	B	D	B	A	C
FEXFEU_01	B	D	B	A	C
FERYOR	B	D	B	E	B
FERYUX	B	D	B	E	B
GOKBOX	B	D	B	E	E
SEBQIA	B	D	B	E	B
WOYQEG	B	D	B	A	B
CAWVOL	B	D	B	C	C
MERWIQ	C	E	F	G	G
EXUCIJ_01	D	B	C	E	D
EXUCIJ_02	D	B	C	E	D
HALLOV	D	B	C	E	D
YASVAQ	D	B	C	E	D
SIHHAS	D	B	D	D	A
FIVGOG	D	C	C	A	C
JKDEM	E	F	E	E	B
QEVPUK	E	F	E	A	B

Appendix B

Triangles

Table B.1: Table describing the variables of 3-aminobutan-2-ol.

Variable	σ	Range	Minimum	Maximum	Mean	Figure
d_1.2	0.019	0.116	1.447	1.563	1.524	
d_1.3	0.055	0.252	2.371	2.623	2.508	
d_1.4	0.417	1.566	2.340	3.906	3.056	
d_1.5	0.039	0.181	2.382	2.563	2.467	
d_1.6	0.395	1.114	2.668	3.782	3.248	
d_2.3	0.038	0.343	1.277	1.620	1.528	
d_2.4	0.074	0.432	2.184	2.616	2.503	
d_2.5	0.017	0.125	1.404	1.529	1.470	
d_2.6	0.049	0.366	2.136	2.502	2.415	
d_3.4	0.036	0.257	1.293	1.550	1.515	

Continued on next page

Table B.1 – continued from previous page

Variable	σ	Range	Minimum	Maximum	Mean	Figure
d.3.5	0.049	0.314	2.240	2.554	2.475	
d.3.6	0.022	0.143	1.302	1.445	1.421	
d.4.5	0.323	1.086	2.793	3.879	3.607	
d.4.6	0.079	0.636	1.863	2.499	2.402	
d.5.6	0.275	1.135	2.631	3.766	2.951	
a.2.1.3	1.976	12.864	26.221	39.085	34.760	
a.2.1.4	14.301	50.435	21.919	72.354	53.534	
a.2.1.5	1.176	5.136	30.928	36.064	33.816	
a.2.1.6	15.790	45.457	18.536	63.993	42.729	
a.3.1.4	7.269	26.116	11.777	37.893	28.374	
a.3.1.5	1.328	7.148	55.815	62.963	59.640	
a.3.1.6	7.612	21.482	10.112	31.594	22.770	
a.4.1.5	14.752	53.918	46.542	100.460	81.700	
a.4.1.6	4.383	15.605	37.632	53.237	43.361	
a.5.1.6	11.172	44.983	45.143	90.126	60.522	
a.1.2.3	3.687	21.603	102.136	123.739	110.589	

Continued on next page

Table B.1 – continued from previous page

Variable	σ	Range	Minimum	Maximum	Mean	Figure
a_1.2.4	21.265	74.552	70.481	145.033	97.764	
a_1.2.5	2.186	9.759	106.933	116.692	110.951	
a_1.2.6	24.090	67.641	81.923	149.564	112.592	
a_3.2.4	1.983	8.791	30.796	39.587	34.435	
a_3.2.5	2.366	9.134	107.954	117.088	111.242	
a_3.2.6	1.053	5.730	30.578	36.308	33.581	
a_4.2.5	20.853	66.801	84.540	151.341	132.324	
a_4.2.6	1.635	11.100	51.079	62.179	58.409	
a_5.2.6	15.851	67.537	82.681	150.218	97.049	
a_1.3.2	1.795	8.905	30.040	38.945	34.651	
a_1.3.4	22.223	79.401	69.454	148.855	98.090	
a_1.3.5	1.390	6.611	56.336	62.947	59.348	
a_1.3.6	24.943	71.891	80.489	152.380	112.034	
a_2.3.4	3.874	18.098	100.139	118.237	110.781	
a_2.3.5	1.388	7.009	30.734	37.743	33.630	
a_2.3.6	2.042	12.017	104.096	116.113	109.913	

















Continued on next page

Table B.1 – continued from previous page

Variable	σ	Range	Minimum	Maximum	Mean	Figure
a_4.3.5	20.993	68.279	85.804	154.083	131.657	
a_4.3.6	3.250	26.180	91.755	117.935	109.772	
a_5.3.6	16.141	66.421	79.323	145.744	95.906	
a_1.4.2	6.991	25.258	13.048	38.306	28.702	
a_1.4.3	14.987	54.308	19.368	73.676	53.536	
a_1.4.5	2.650	13.742	38.471	52.213	41.080	
a_1.4.6	15.172	55.011	44.547	99.558	72.534	
a_2.4.3	1.988	9.307	30.967	40.274	34.784	
a_2.4.5	6.664	20.767	10.885	31.652	16.933	
a_2.4.6	1.419	7.679	55.457	63.136	58.953	
a_3.4.5	13.981	44.962	16.489	61.451	30.661	
a_3.4.6	1.801	15.207	29.117	44.324	33.837	
a_5.4.6	11.622	47.421	43.496	90.917	54.657	
a_1.5.2	1.078	4.624	32.379	37.003	35.233	
a_1.5.3	1.809	8.269	56.349	64.618	61.012	
a_1.5.4	14.950	54.699	38.366	93.065	57.220	

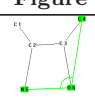
Continued on next page

Table B.1 – continued from previous page

Variable	σ	Range	Minimum	Maximum	Mean	Figure
a_1.5.6	13.202	44.592	48.331	92.923	73.639	
a_2.5.3	1.249	4.765	32.138	36.903	35.129	
a_2.5.4	14.204	46.035	17.773	63.808	30.743	
a_2.5.6	10.608	44.127	18.681	62.808	53.829	
a_3.5.4	7.026	23.317	9.428	32.745	17.682	
a_3.5.6	5.009	20.178	12.474	32.652	28.050	
a_4.5.6	3.113	16.770	32.608	49.378	40.562	
a_1.6.2	8.319	23.114	11.899	35.013	24.679	
a_1.6.3	17.367	50.543	17.508	68.051	45.196	
a_1.6.4	15.527	57.119	39.093	96.212	64.105	
a_1.6.5	4.492	12.307	39.595	51.902	45.840	
a_2.6.3	1.167	6.404	33.309	39.713	36.506	
a_2.6.4	1.952	9.750	56.221	65.971	62.639	
a_2.6.5	5.293	23.646	11.102	34.748	29.122	
a_3.6.4	1.521	10.974	32.948	43.922	36.391	
a_3.6.5	11.173	46.243	21.782	68.025	56.045	

Continued on next page

Table B.1 – continued from previous page

Variable	σ	Range	Minimum	Maximum	Mean	Figure
a_4.6.5	12.867	48.596	48.981	97.577	84.780	

Appendix C

Factor Analysis

C.1 3-chlorobut-2-ene-thiolate

Table C.1: Table of rotated factor loadings of the fragment 3-chlorobut-2-ene-thiolate

Variable	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6
d.1.2	-0.098	-0.180	-0.126	-0.855	-0.008	0.103
d.1.3	0.191	0.150	-0.022	0.841	0.289	-0.033
d.1.4	0.759	0.305	-0.074	-0.176	-0.524	0.041
d.1.5	0.017	0.638	-0.097	-0.014	0.258	0.163
d.1.6	-0.745	-0.042	-0.386	0.166	0.496	0.048
d.2.3	0.140	0.937	-0.073	0.210	0.028	-0.053
d.2.4	0.422	0.073	0.117	0.759	0.046	0.111
d.2.5	0.017	-0.777	0.055	-0.248	0.144	0.516
d.2.6	0.352	0.644	0.045	0.416	-0.155	0.268
d.3.4	0.902	0.354	0.023	0.001	-0.240	-0.027
d.3.5	-0.065	-0.696	-0.081	0.430	0.383	-0.279
d.3.6	-0.952	0.131	-0.152	0.050	0.212	-0.017
d.4.5	-0.955	-0.167	-0.097	-0.173	-0.103	0.070
d.4.6	0.375	0.083	0.836	0.340	-0.00696	0.119
d.5.6	0.952	-0.110	-0.017	0.212	0.141	0.106
a.2.1.3	0.076	0.990	-0.043	0.034	-0.081	-0.044
a.2.1.4	-0.691	-0.288	0.097	0.313	0.569	-0.012
a.2.1.5	0.047	-0.890	0.145	0.158	0.045	0.382
a.2.1.6	0.737	0.225	0.357	-0.150	-0.500	0.020
a.3.1.4	0.924	0.352	0.049	0.046	-0.123	-0.019
a.3.1.5	-0.189	-0.910	-0.028	-0.214	0.077	-0.261
a.3.1.6	-0.971	0.1919	-0.088	-0.070	0.059	-0.047
a.4.1.5	-0.954	-0.261	-0.045	-0.024	0.117	0.052
a.4.1.6	0.3451	-0.030	0.908	0.221	-0.062	0.017
a.5.1.6	0.981	-0.093	0.093	0.091	-0.070	0.073

Continued on next page

Table C.1 – continued from previous page [3-chlorobut-2-ene-thiolate]

Variable	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6
a_1.2.3	-0.054	-0.972	0.054	0.171	0.118	0.032
a_1.2.4	0.728	0.301	-0.080	-0.211	-0.569	0.012
a_1.2.5	-0.027	0.907	-0.126	-0.043	-0.016	-0.358
a_1.2.6	-0.728	-0.196	-0.356	0.211	0.505	-0.011
a_3.2.4	0.953	0.147	0.051	-0.033	-0.257	-0.008
a_3.2.5	-0.157	-0.827	-0.026	0.284	0.200	-0.381
a_3.2.6	-0.948	-0.122	-0.153	0.034	0.248	0.008
a_4.2.5	-0.956	0.012	-0.129	-0.215	-0.144	-0.020
a_4.2.6	0.137	-0.186	0.969	-0.051	0.037	-0.009
a_5.2.6	0.962	0.012	-0.038	0.226	0.142	0.014
a_1.3.2	-0.094	-0.962	0.029	-0.241	0.040	0.054
a_1.3.4	-0.925	-0.346	-0.062	-0.069	0.118	0.018
a_1.3.5	0.138	0.930	0.028	0.004	-0.131	0.287
a_1.3.6	0.969	-0.2037	0.089	0.050	-0.069	0.050
a_2.3.4	-0.949	-0.153	-0.060	0.040	0.260	0.008
a_2.3.5	-0.025	-0.732	0.112	-0.467	-0.099	0.460
a_2.3.6	0.943	0.117	0.160	-0.042	-0.262	-0.006
a_4.3.5	-0.951	-0.268	-0.060	-0.074	0.107	0.047
a_4.3.6	0.767	-0.049	0.588	-0.067	-0.226	-0.003
a_5.3.6	0.979	-0.121	0.092	0.049	-0.082	0.074
a_1.4.2	-0.758	-0.310	0.059	0.092	0.558	-0.011
a_1.4.3	-0.920	-0.364	-0.022	0.006	0.134	0.019
a_1.4.5	0.950	0.281	0.046	0.019	-0.107	-0.06
a_1.4.6	-0.751	-0.152	-0.435	0.071	0.462	-0.002
a_2.4.3	-0.954	-0.144	-0.047	0.030	0.255	0.008
a_2.4.5	0.960	-0.015	0.128	0.197	0.142	0.021
a_2.4.6	-0.160	0.242	-0.950	-0.015	-0.063	0.021
a_3.4.5	0.634	-0.584	0.242	0.157	0.153	-0.323
a_3.4.6	-0.952	-0.119	-0.134	0.027	0.245	0.010
a_5.4.6	0.969	0.017	-0.008	0.199	0.139	0.021
a_1.5.2	-0.074	0.823	-0.166	-0.326	-0.086	-0.401
a_1.5.3	0.230	0.812	0.026	0.435	-0.008	0.210
a_1.5.4	0.957	0.250	0.044	0.028	-0.123	-0.050
a_1.5.6	-0.981	0.086	-0.097	-0.093	0.070	-0.070
a_2.5.3	0.112	0.970	-0.055	0.123	-0.059	-0.059
a_2.5.4	0.945	-0.008	0.131	0.252	0.147	0.017
a_2.5.6	-0.961	-0.002	0.039	-0.214	-0.161	-0.008
a_3.5.4	0.937	0.314	0.046	0.066	-0.120	-0.028
a_3.5.6	-0.977	0.152	-0.094	-0.059	0.067	-0.056
a_4.5.6	-0.426	-0.093	0.876	0.097	-0.124	0.006
a_1.6.2	0.708	0.152	0.352	-0.297	-0.507	-0.004
a_1.6.3	0.972	-0.166	0.086	0.120	-0.034	0.042
a_1.6.4	0.735	0.225	-0.116	-0.263	-0.566	-0.010
a_1.6.5	-0.980	0.107	-0.085	-0.088	0.072	-0.077
a_2.6.3	0.950	0.125	0.148	-0.029	-0.239	-0.009
a_2.6.4	-0.103	0.108	-0.967	0.136	-0.004	-0.007
a_2.6.5	-0.961	-0.016	0.038	-0.233	-0.132	-0.017

Continued on next page

Table C.1 – continued from previous page [3-chlorobut-2-ene-thiolate]

Variable	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6
a_3.6.4	0.956	0.137	0.075	-0.022	-0.243	-0.010
a_3.6.5	-0.879	-0.245	-0.058	0.078	0.239	-0.270
a_4.6.5	-0.963	-0.009	-0.078	-0.216	-0.132	-0.023

Table C.2: Table of communalities of the fragment 3-chlorobut-2-ene-thiolate

Variable	Initial	Extraction
Communalities	Initial	Extraction
d.1.2	1	0.799
d.1.3	1	0.857
d.1.4	1	0.981
d.1.5	1	0.510
d.1.6	1	0.981
d.2.3	1	0.951
d.2.4	1	0.788
d.2.5	1	0.956
d.2.6	1	0.810
d.3.4	1	0.997
d.3.5	1	0.904
d.3.6	1	0.995
d.4.5	1	0.994
d.4.6	1	0.976
d.5.6	1	0.996
a.2.1.3	1	0.998
a.2.1.4	1	0.991
a.2.1.5	1	0.987
a.2.1.6	1	0.994
a.3.1.4	1	0.998
a.3.1.5	1	0.985
a.3.1.6	1	0.998
a.4.1.5	1	0.998
a.4.1.6	1	0.997
a.5.1.6	1	0.998
a.1.2.3	1	0.996
a.1.2.4	1	0.995
a.1.2.5	1	0.969
a.1.2.6	1	0.995
a.3.2.4	1	0.999
a.3.2.5	1	0.974
a.3.2.6	1	0.999
a.4.2.5	1	0.998
a.4.2.6	1	0.996
a.5.2.6	1	0.998
Continued on next page		

Table C.2 – continued from previous page [pentanone-2-one]

Variable	Initial	Extraction
a.1.3.2	1	0.997
a.1.3.4	1	0.999
a.1.3.5	1	0.984
a.1.3.6	1	0.9988
a.2.3.4	1	0.998
a.2.3.5	1	0.988
a.2.3.6	1	0.999
a.4.3.5	1	0.998
a.4.3.6	1	0.992
a.5.3.6	1	0.997
a.1.4.2	1	0.994
a.1.4.3	1	0.996
a.1.4.5	1	0.998
a.1.4.6	1	0.995
a.2.4.3	1	0.999
a.2.4.5	1	0.998
a.2.4.6	1	0.991
a.3.4.5	1	0.954
a.3.4.6	1	0.999
a.5.4.6	1	0.998
a.1.5.2	1	0.985
a.1.5.3	1	0.946
a.1.5.4	1	0.998
a.1.5.6	1	0.998
a.2.5.3	1	0.979
a.2.5.4	1	0.995
a.2.5.6	1	0.997
a.3.5.4	1	0.998
a.3.5.6	1	0.997
a.4.5.6	1	0.982
a.1.6.2	1	0.994
a.1.6.3	1	0.996
a.1.6.4	1	0.994
a.1.6.5	1	0.998
a.2.6.3	1	0.999
a.2.6.4	1	0.977
a.2.6.5	1	0.998
a.3.6.4	1	0.999
a.3.6.5	1	0.972
a.4.6.5	1	0.997

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	43.971	58.628	58.628	43.971	58.628	58.628	41.010	54.681	54.681
2	14.492	19.322	77.951	14.492	19.322	77.951	14.830	19.774	74.454
3	6.097	8.129	86.080	6.097	8.129	86.080	6.655	8.874	83.328
4	5.592	7.457	93.536	5.592	7.457	93.536	4.460	5.947	89.275
5	1.668	2.223	95.760	1.668	2.223	95.760	4.334	5.779	95.054
6	1.195	1.593	97.353	1.195	1.593	97.353	1.725	2.300	97.352

Table C.3: Table of variances of factor analysis of the fragment 3-chlorobut-2-ene-thiolate

C.2 3-aminobutan-2-ol

Table C.4: Table of rotated factor loadings of the fragment 3-aminobutan-2-ol

Variable	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6
d.1.2	-0.114	0.206	-0.127	0.638	0.024	0.200
d.1.3	-0.102	0.116	-0.155	0.290	0.746	-0.177
d.1.4	-0.830	0.233	-0.145	0.050	0.303	-0.262
d.1.5	-0.002	0.107	-0.033	0.305	-0.235	0.812
d.1.6	0.144	-0.972	0.120	0.106	0.007	0.006
d.2.3	-0.176	0.047	-0.140	0.813	0.043	0.081
d.2.4	-0.217	-0.010	0.023	0.520	0.765	-0.050
d.2.5	0.059	-0.189	0.179	0.188	0.023	-0.023
d.2.6	-0.095	-0.025	-0.188	0.709	-0.038	0.243
d.3.4	0.028	-0.116	0.114	0.757	0.038	0.175
d.3.5	-0.113	0.034	-0.082	0.554	-0.140	0.064
d.3.6	0.205	-0.159	0.051	0.572	0.128	0.147
d.4.5	0.869	-0.029	0.298	0.077	0.099	-0.206
d.4.6	0.040	-0.122	0.051	0.957	-0.075	0.019
d.5.6	-0.302	0.146	-0.925	0.088	0.007	0.011
a.2.1.3	-0.064	-0.050	0.013	0.497	-0.635	0.252
a.2.1.4	0.879	-0.261	0.175	0.053	-0.105	0.243
a.2.1.5	0.011	-0.184	0.103	-0.080	0.278	-0.859
a.2.1.6	-0.142	0.972	-0.145	-0.001	-0.012	0.053
a.3.1.4	0.888	-0.271	0.171	0.039	-0.137	0.210
a.3.1.5	-0.055	-0.074	0.017	0.269	-0.460	-0.149
a.3.1.6	-0.107	0.975	-0.117	-0.067	0.055	0.035
a.4.1.5	0.945	-0.135	0.250	0.007	-0.143	-0.007
a.4.1.6	0.688	0.600	0.111	0.167	-0.134	0.219
a.5.1.6	-0.295	0.672	-0.671	-0.016	0.000	-0.017
a.1.2.3	0.038	0.041	-0.033	-0.422	0.688	-0.279
a.1.2.4	-0.882	0.255	-0.174	-0.036	0.118	-0.235
a.1.2.5	0.030	0.115	-0.047	-0.025	-0.308	0.898
a.1.2.6	0.137	-0.975	0.140	0.012	0.010	-0.055
a.3.2.4	0.226	-0.070	0.014	0.170	-0.870	0.210
a.3.2.5	-0.025	0.079	-0.057	-0.020	-0.247	0.031
a.3.2.6	0.223	-0.061	0.247	-0.221	0.184	-0.209
a.4.2.5	0.890	-0.004	0.289	-0.046	-0.109	-0.199
a.4.2.6	0.216	-0.149	0.121	0.706	-0.549	-0.024
a.5.2.6	-0.317	0.148	-0.933	-0.021	-0.005	-0.006
a.1.3.2	-0.008	-0.029	0.054	0.318	-0.714	0.295
a.1.3.4	-0.891	0.253	-0.152	-0.006	0.137	-0.219
a.1.3.5	0.101	0.001	0.093	-0.183	-0.502	0.665
a.1.3.6	0.130	-0.972	0.131	0.065	-0.090	0.001
a.2.3.4	-0.177	0.032	0.053	-0.198	0.896	-0.192
a.2.3.5	0.092	-0.112	0.122	-0.271	0.191	-0.064
a.2.3.6	-0.088	-0.008	-0.187	-0.024	-0.182	0.231
a.4.3.5	0.822	-0.011	0.289	-0.111	0.148	-0.248

Continued on next page

Table C.4 – continued from previous page [3-aminobutan-2-ol]

Variable	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6
a_4.3.6	-0.038	-0.071	-0.002	0.926	-0.200	-0.131
a_5.3.6	-0.313	0.137	-0.932	-0.021	0.010	0.018
a_1.4.2	0.886	-0.242	0.171	0.002	-0.145	0.218
a_1.4.3	0.891	-0.243	0.142	-0.010	-0.136	0.222
a_1.4.5	0.020	-0.296	-0.126	-0.061	-0.013	0.621
a_1.4.6	0.507	-0.810	0.146	0.009	-0.154	0.132
a_2.4.3	0.120	0.008	-0.118	0.217	-0.877	0.164
a_2.4.5	-0.884	-0.005	-0.290	0.029	0.092	0.205
a_2.4.6	0.038	0.077	-0.230	-0.420	-0.537	0.238
a_3.4.5	-0.825	0.015	-0.295	0.103	-0.148	0.239
a_3.4.6	0.058	0.070	-0.018	-0.942	0.183	0.083
a_5.4.6	-0.628	0.104	-0.734	0.004	-0.051	0.115
a_1.5.2	-0.073	-0.032	-0.018	0.138	0.320	-0.884
a_1.5.3	-0.038	0.053	-0.084	-0.057	0.723	-0.401
a_1.5.4	-0.936	0.185	-0.224	0.004	0.144	-0.103
a_1.5.6	0.218	-0.869	0.432	0.041	0.013	-0.043
a_2.5.3	-0.054	-0.026	-0.028	0.339	0.256	0.013
a_2.5.4	-0.892	0.009	-0.288	0.054	0.117	0.196
a_2.5.6	0.314	-0.149	0.934	0.050	0.002	0.025
a_3.5.4	-0.815	0.003	-0.278	0.125	-0.149	0.265
a_3.5.6	0.339	-0.139	0.921	0.007	0.022	-0.029
a_4.5.6	-0.695	-0.099	0.327	0.387	-0.105	0.212
a_1.6.2	-0.127	0.977	-0.129	-0.033	-0.007	0.058
a_1.6.3	-0.140	0.968	-0.137	-0.064	0.105	-0.017
a_1.6.4	-0.689	0.622	-0.174	-0.056	0.188	-0.191
a_1.6.5	0.094	0.882	0.400	-0.080	-0.040	0.168
a_2.6.3	-0.047	0.070	0.104	0.241	0.151	-0.214
a_2.6.4	-0.208	0.069	0.066	-0.286	0.850	-0.153
a_2.6.5	0.321	-0.146	0.923	-0.039	0.011	-0.032
a_3.6.4	0.011	0.069	0.024	-0.863	0.212	0.181
a_3.6.5	0.300	-0.135	0.934	0.027	-0.024	-0.014
a_4.6.5	0.735	-0.070	0.584	-0.097	0.071	-0.155

Table C.5: Table of communalities of the fragment 3-aminobutan-2-ol

Variable	Initial	Extraction
Communalities	Initial	Extraction
d_1.2	1	0.554
d_1.3	1	0.752
d_1.4	1	0.964
d_1.5	1	0.786
d_1.6	1	0.987
d_2.3	1	0.793
Continued on next page		

Table C.5 – continued from previous page [3-aminobutan-2-ol]

Variable	Initial	Extraction
d.2.4	1	0.770
d.2.5	1	0.202
d.2.6	1	0.734
d.3.4	1	0.700
d.3.5	1	0.750
d.3.6	1	0.521
d.4.5	1	0.972
d.4.6	1	0.969
d.5.6	1	0.972
a.2.1.3	1	0.820
a.2.1.4	1	0.929
a.2.1.5	1	0.814
a.2.1.6	1	0.988
a.3.1.4	1	0.953
a.3.1.5	1	0.915
a.3.1.6	1	0.991
a.4.1.5	1	0.995
a.4.1.6	1	0.930
a.5.1.6	1	0.982
a.1.2.3	1	0.844
a.1.2.4	1	0.935
a.1.2.5	1	0.817
a.1.2.6	1	0.988
a.3.2.4	1	0.816
a.3.2.5	1	0.633
a.3.2.6	1	0.592
a.4.2.5	1	0.967
a.4.2.6	1	0.772
a.5.2.6	1	0.959
a.1.3.2	1	0.826
a.1.3.4	1	0.949
a.1.3.5	1	0.788
a.1.3.6	1	0.994
a.2.3.4	1	0.809
a.2.3.5	1	0.663
a.2.3.6	1	0.575
a.4.3.5	1	0.969
a.4.3.6	1	0.795
a.5.3.6	1	0.965
a.1.4.2	1	0.943
a.1.4.3	1	0.945
a.1.4.5	1	0.677
a.1.4.6	1	0.992
a.2.4.3	1	0.751
a.2.4.5	1	0.963
a.2.4.6	1	0.782
a.3.4.5	1	0.969

Continued on next page

Table C.5 – continued from previous page [3-aminobutan-2-ol]

Variable	Initial	Extraction
a_3.4.6	1	0.841
a_5.4.6	1	0.989
a_1.5.2	1	0.772
a_1.5.3	1	0.897
a_1.5.4	1	0.987
a_1.5.6	1	0.987
a_2.5.3	1	0.580
a_2.5.4	1	0.968
a_2.5.6	1	0.948
a_3.5.4	1	0.965
a_3.5.6	1	0.963
a_4.5.6	1	0.879
a_1.6.2	1	0.985
a_1.6.3	1	0.994
a_1.6.4	1	0.981
a_1.6.5	1	0.973
a_2.6.3	1	0.547
a_2.6.4	1	0.784
a_2.6.5	1	0.973
a_3.6.4	1	0.680
a_3.6.5	1	0.962
a_4.6.5	1	0.986

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	23.178	30.904	30.904	23.178	30.904	30.904	16.680	22.240	22.240
2	14.625	19.500	50.404	14.625	19.500	50.404	11.134	14.846	37.085
3	9.563	12.750	63.154	9.563	12.750	63.154	9.654	12.872	49.958
4	7.986	10.647	73.802	7.986	10.647	73.802	9.238	12.317	62.275
5	4.932	6.576	80.378	4.932	6.576	80.378	8.972	11.964	74.239
6	3.775139	5.034	85.411	3.775	5.034	85.411	8.379	11.172	85.411

Table C.6: Table of variances of factor analysis of the fragment 3-aminobutan-2-ol

C.3 Pentanone-2-one

Table C.7: Table of rotated factor loadings of the fragment pentanone-2-one

Variable	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6
d.1.2	-0.008	-0.182	0.804	-0.007	0.064	0.444
d.1.3	0.127	-0.116	-0.373	0.343	0.122	0.717
d.1.4	-0.028	-0.948	0.232	-0.163	0.056	0.047
d.1.5	-0.875	-0.436	0.112	-0.001	-0.070	0.039
d.1.6	-0.025	-0.070	0.757	0.042	0.055	0.384
d.2.3	0.165	-0.177	0.234	0.682	0.116	0.103
d.2.4	0.135	-0.322	0.358	-0.506	0.352	-0.028
d.2.5	-0.991	-0.072	0.056	-0.014	-0.044	-0.002
d.2.6	0.094	0.202	-0.837	0.060	0.083	-0.124
d.3.4	0.148	0.127	-0.058	0.495	0.434	-0.039
d.3.5	-0.035	0.035	0.099	0.151	-0.804	-0.030
d.3.6	0.161	-0.093	0.526	0.288	0.199	-0.511
d.4.5	-0.125	0.141	0.193	-0.030	-0.180	-0.039
d.4.6	0.077	0.981	-0.020	-0.023	0.086	-0.042
d.5.6	-0.981	0.154	0.0419	-0.040	0.017	-0.012
a.2.1.3	-0.059	-0.072	0.864	-0.026	-0.033	-0.340
a.2.1.4	0.0430	0.989	-0.051	0.090	0.008	-0.014
a.2.1.5	-0.410	0.815	0.011	-0.069	0.088	-0.052
a.2.1.6	0.108	0.125	-0.917	0.004	0.072	-0.165
a.3.1.4	0.063	0.890	-0.336	0.290	-0.013	0.051
a.3.1.5	0.880	0.410	-0.155	0.074	-0.012	0.030
a.3.1.6	0.036	0.037	0.309	-0.040	0.037	-0.936
a.4.1.5	0.956	0.006	0.063	-0.092	0.115	-0.032
a.4.1.6	0.054	0.984	-0.118	0.069	0.018	-0.046
a.5.1.6	-0.740	0.612	-0.026	-0.062	0.088	-0.052
a.1.2.3	0.068	0.088	-0.941	0.090	0.026	0.259
a.1.2.4	-0.041	-0.991	0.034	-0.091	-0.005	0.005
a.1.2.5	0.316	-0.834	-0.028	0.077	-0.095	0.041
a.1.2.6	-0.125	-0.011	0.672	0.011	-0.114	-0.035
a.3.2.4	-0.004	0.304	-0.280	0.869	-0.039	0.031
a.3.2.5	0.992	0.069	-0.044	0.056	-0.053	0.002
a.3.2.6	-0.019	-0.123	0.878	-0.135	0.039	-0.358
a.4.2.5	0.993	0.053	-0.012	-0.028	0.090	-0.003
a.4.2.6	0.057	0.990	-0.022	0.049	0.024	-0.021
a.5.2.6	-0.845	0.476	0.042	-0.069	0.098	-0.024
a.1.3.2	-0.074	-0.098	0.969	-0.130	-0.019	-0.140
a.1.3.4	-0.062	-0.891	0.335	-0.281	0.019	-0.073
a.1.3.5	-0.880	-0.408	0.158	-0.076	0.009	-0.034
a.1.3.6	-0.146	0.036	0.777	-0.257	-0.082	0.0367
a.2.3.4	-0.001	-0.257	0.237	-0.912	0.091	-0.054
a.2.3.5	-0.993	-0.065	0.043	-0.061	0.047	-0.005
a.2.3.6	0.019	0.169	-0.923	0.039	-0.023	0.232
a.4.3.5	-0.035	0.111	0.044	-0.104	0.906	-0.011

Continued on next page

Table C.7 – continued from previous page [pentanone-2-one]

Variable	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6
a_4.3.6	0.057	0.987	-0.074	-0.057	0.040	0.016
a_5.3.6	-0.973	0.202	0.0104	-0.047	0.068	-0.002
a_1.4.2	0.038	0.991	-0.001	0.093	-0.003	0.012
a_1.4.3	0.062	0.892	-0.332	0.277	-0.022	0.084
a_1.4.5	-0.959	0.099	-0.054	0.106	-0.122	0.023
a_1.4.6	0.008	0.948	-0.024	0.149	-0.056	-0.010
a_2.4.3	0.006	0.202	-0.184	0.927	-0.140	0.075
a_2.4.5	-0.993	-0.049	0.0142	0.030	-0.090	0.002
a_2.4.6	-0.053	-0.970	-0.130	-0.012	-0.039	-0.015
a_3.4.5	-0.020	-0.091	0.0160	0.021	-0.981	0.004
a_3.4.6	-0.055	-0.985	0.092	0.056	-0.050	-0.033
a_5.4.6	-0.975	-0.172	0.035	-0.021	-0.029	-0.004
a_1.5.2	-0.025	0.826	0.076	-0.094	0.105	-0.007
a_1.5.3	0.880	0.406	-0.160	0.077	-0.006	0.038
a_1.5.4	0.956	-0.129	0.051	-0.109	0.124	-0.021
a_1.5.6	0.652	0.542	0.024	-0.079	0.098	-0.007
a_2.5.3	0.992	0.059	-0.042	0.071	-0.035	0.009
a_2.5.4	0.993	0.047	-0.016	-0.031	0.090	-0.001
a_2.5.6	0.963	-0.216	-0.120	0.029	-0.023	-0.006
a_3.5.4	0.066	0.067	-0.066	0.052	0.972	0.003
a_3.5.6	0.974	-0.206	-0.004	0.048	-0.048	-0.003
a_4.5.6	0.973	0.190	-0.037	0.025	0.027	0.005
a_1.6.2	0.088	-0.144	-0.034	-0.027	0.116	0.281
a_1.6.3	0.069	-0.046	-0.686	0.187	0.028	0.578
a_1.6.4	-0.046	-0.986	0.102	-0.085	-0.004	0.040
a_1.6.5	0.558	-0.731	0.019	0.080	-0.110	0.053
a_2.6.3	0.018	0.061	-0.774	0.238	-0.055	0.461
a_2.6.4	-0.057	-0.989	0.066	-0.059	-0.019	0.032
a_2.6.5	0.785	-0.547	-0.016	0.079	-0.119	0.032
a_3.6.4	-0.060	-0.985	0.036	0.057	-0.016	0.022
a_3.6.5	0.972	-0.199	-0.016	0.047	-0.088	0.006
a_4.6.5	0.973	0.131	-0.031	0.015	0.034	0.004

Table C.8: Table of commonalities of the fragment pentanone-2-one

Variable	Initial	Extraction
d_1.2	1	0.978
d_1.3	1	0.961
d_1.4	1	0.993
d_1.5	1	0.994
d_1.6	1	0.859
d_2.3	1	0.781
d_2.4	1	0.848
d_2.5	1	0.999

Continued on next page

Table C.8 – continued from previous page [pentanone-2-one]

Variable	Initial	Extraction
d.2.6	1	0.831
d.3.4	1	0.626
d.3.5	1	0.976
d.3.6	1	0.835
d.4.5	1	0.778
d.4.6	1	0.983
d.5.6	1	0.999
a.2.1.3	1	0.988
a.2.1.4	1	0.996
a.2.1.5	1	0.986
a.2.1.6	1	0.981
a.3.1.4	1	0.996
a.3.1.5	1	0.974
a.3.1.6	1	0.993
a.4.1.5	1	0.969
a.4.1.6	1	0.993
a.5.1.6	1	0.994
a.1.2.3	1	0.998
a.1.2.4	1	0.997
a.1.2.5	1	0.985
a.1.2.6	1	0.986
a.3.2.4	1	0.942
a.3.2.5	1	0.998
a.3.2.6	1	0.999
a.4.2.5	1	0.998
a.4.2.6	1	0.991
a.5.2.6	1	0.993
a.1.3.2	1	0.997
a.1.3.4	1	0.996
a.1.3.5	1	0.974
a.1.3.6	1	0.930
a.2.3.4	1	0.977
a.2.3.5	1	0.998
a.2.3.6	1	0.987
a.4.3.5	1	0.878
a.4.3.6	1	0.989
a.5.3.6	1	0.996
a.1.4.2	1	0.997
a.1.4.3	1	0.995
a.1.4.5	1	0.987
a.1.4.6	1	0.962
a.2.4.3	1	0.970
a.2.4.5	1	0.999
a.2.4.6	1	0.967
a.3.4.5	1	0.975
a.3.4.6	1	0.989
a.5.4.6	1	0.990

Continued on next page

Table C.8 – continued from previous page [pentanone-2-one]

Variable	Initial	Extraction
a_1.5.2	1	0.977
a_1.5.3	1	0.974
a_1.5.4	1	0.989
a_1.5.6	1	0.962
a_2.5.3	1	0.998
a_2.5.4	1	0.999
a_2.5.6	1	0.992
a_3.5.4	1	0.994
a_3.5.6	1	0.996
a_4.5.6	1	0.994
a_1.6.2	1	0.996
a_1.6.3	1	0.983
a_1.6.4	1	0.997
a_1.6.5	1	0.992
a_2.6.3	1	0.970
a_2.6.4	1	0.994
a_2.6.5	1	0.992
a_3.6.4	1	0.982
a_3.6.5	1	0.996
a_4.6.5	1	0.976

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	23.911	31.881	31.881	23.911	31.881	31.881	23.448	31.264	31.264
2	22.244	29.659	61.540	22.244	29.659	61.540	21.231	28.307	59.571
3	10.931	14.575	76.115	10.931	14.575	76.115	10.587	14.116	73.687
4	4.588	6.117	82.232	4.588	6.117	82.232	4.326	5.768	79.455
5	3.495	4.659	86.891	3.495	4.660	86.891	4.080	5.440	84.895
6	2.703	3.603	90.494	2.703	3.603	90.494	3.156	4.208	89.103
7	2.119	2.826	93.320	2.119	2.826	93.320	2.657	3.542	92.645
8	1.474	1.965	95.286	1.474	1.965	95.286	1.531	2.041	94.686
9	1.035	1.380	96.666	1.036	1.380	96.666	1.485	1.980	96.663

Table C.9: Table of variances of factor analysis of the fragment pentanone

Appendix D

Sparse principal components analysis

D.1 Difluoroalkene

Table D.1: First Eigenvalue of DSPCA of the fragment difluoroalkene.

Variable	Value
d12	0.000
d13	0.000
d14	0.000
d15	0.000
d16	0.000
d23	0.000
d24	0.000
d25	0.000
d26	0.000
d34	-0.002
d35	0.000
d36	0.001
d45	0.002
d46	0.000
d56	-0.002
a213	-0.021
a214	0.009
a215	0.010
a216	-0.004
a314	-0.294
a315	0.007
a316	0.222
Continued ..	

Variable	Value
a415	0.274
a416	0.005
a516	-0.238
a123	0.010
a124	-0.019
a125	-0.005
a126	0.007
a324	-0.285
a325	0.004
a326	0.261
a425	0.232
a426	0.009
a526	-0.252
a132	0.008
a134	0.183
a135	-0.004
a136	-0.132
a234	0.076
a235	0.004
a236	-0.058
a435	0.173
a436	0.007
a536	-0.141
a142	0.007
a143	0.080
a145	-0.061
a146	0.002
a243	0.179
a245	-0.138
a246	-0.005
a345	0.007
a346	0.167
a546	-0.149
a152	-0.005
a153	-0.003
a154	-0.183
a156	0.153
a253	-0.008
a254	-0.064
a256	0.060
a354	-0.192
a356	0.146
a456	-0.002
a162	-0.003
a163	-0.061
a164	-0.007
a165	0.055
a263	-0.174
Continued ..	

Variable	Value
a264	-0.004
a265	0.162
a364	-0.185
a365	-0.002
a465	0.153

D.2 3-aminobutan-2-ol

Table D.2: First Eigenvector of DSPCA of the fragment 3aminobutan-2-ol.

Variable	Value
d12	0.000
d13	0.000
d14	-0.003
d15	0.000
d16	0.002
d23	0.000
d24	0.000
d25	0.000
d26	0.000
d34	0.000
d35	0.000
d36	0.000
d45	0.002
d46	0.000
d56	-0.001
a213	0.001
a214	0.194
a215	0.001
a216	-0.153
a314	0.082
a315	0.000
a316	-0.054
a415	0.204
a416	0.009
a516	-0.129
a123	-0.002
a124	-0.309
a125	0.000
a126	0.253
a324	0.004
a325	-0.001
a326	0.002
Continued ..	

Variable	Value
a425	0.267
a426	0.004
a526	-0.154
a132	0.001
a134	-0.324
a135	0.002
a136	0.263
a234	-0.006
a235	0.002
a236	-0.002
a435	0.243
a436	0.001
a536	-0.155
a142	0.077
a143	0.202
a145	0.003
a146	0.205
a243	0.002
a245	-0.061
a246	-0.001
a345	-0.150
a346	0.000
a546	-0.132
a152	-0.001
a153	-0.002
a154	-0.211
a156	0.148
a253	0.000
a254	-0.170
a256	0.091
a354	-0.057
a356	0.032
a456	-0.009
a162	-0.063
a163	-0.174
a164	-0.227
a165	-0.008
a263	0.000
a264	-0.004
a265	0.034
a364	0.000
a365	0.094
a465	0.152

D.3 Pentan-2-one

Table D.3: First Eigenvector of DSPCA of the fragment pentan-2-one.

Variable	Value
d12	0.000
d13	0.000
d14	0.000
d15	-0.002
d16	0.000
d23	0.000
d24	0.000
d25	-0.002
d26	0.000
d34	0.000
d35	0.000
d36	0.000
d45	0.000
d46	0.000
d56	-0.002
a213	0.000
a214	0.008
a215	-0.011
a216	0.000
a314	0.004
a315	0.154
a316	0.000
a415	0.027
a416	0.006
a516	-0.049
a123	0.000
a124	-0.015
a125	0.009
a126	-0.001
a324	0.001
a325	0.185
a326	0.000
a425	0.075
a426	0.009
a526	-0.097
a132	0.000
a134	-0.016
a135	-0.346
a136	-0.001
a234	-0.001
a235	-0.297
a236	0.000
a435	0.000
Continued ..	

Variable	Value
a436	0.007
a536	-0.241
a142	0.003
a143	0.008
a145	-0.203
a146	0.001
a243	0.000
a245	-0.289
a246	-0.002
a345	-0.001
a346	-0.004
a546	-0.523
a152	0.001
a153	0.158
a154	0.151
a156	0.008
a253	0.079
a254	0.182
a256	0.016
a354	0.001
a356	0.098
a456	0.354
a162	0.000
a163	0.000
a164	-0.008
a165	0.030
a263	0.000
a264	-0.006
a265	0.061
a364	-0.002
a365	0.111
a465	0.134

Bibliography

- [1] F. H. Allen. The cambridge structural database: a quarter of a million crystal structures and rising. *Acta Crystallographica Section B*, B58:380–388, 2002.
- [2] F. H. Allen and M. J. Doyle. Automated conformational analysis from crystallographic data. 2. symmetry-modified jarvis-patrick and complete-linkage clustering algorithms for three-dimensional pattern recognition. *Acta Crystallographica*, B47:41–49, 1991.
- [3] F. H. Allen and M. J. Doyle. Automated conformational analysis from crystallographic data. 3. three-dimensional pattern recognition within the cambridge structural database system: Implementation and practical examples. *Acta Crystallographica*, B47:50–61, 1991.
- [4] F. H. Allen and M. J. Doyle. Automated conformational analysis from crystallographic data. i. a symmetry-modified single-linkage clustering algorithm for three-dimensional pattern recognition. *Acta Crystallographica Section A*, B47:29–40, 1991.
- [5] F. H. Allen, S. Harris, and R. Tylor. Comparison of conformer distributions in the crystalline state with conformational energies calculated by ab initio techniques. *Journal of Computer-Aided Molecular Design*, 10:247–254, 1996.
- [6] F. H. Allen and O. Johnson. Automated conformational analysis from crystallographic data. 4. statistical descriptors for a distribution of torsion angles. *Acta Crystallographica*, B47:62–67, 1991.

- [7] G. Barr, W. Dong, C. Gilmore, and J. Faber. High-throughput powder diffraction. iii. the application of full-profile pattern matching and multivariate statistical analysis to round-robin-type data sets. *Journal of Applied Crystallography*, 37(4):635–642, 2004.
- [8] G. Barr, W. Dong, and C. J. Gilmore. High-throughput powder diffraction. ii. applications of clustering methods and multivariate data analysis. *Journal of Applied Crystallography*, 37(2):243–252, 2004.
- [9] G. Barr, W. Dong, and C. J. Gilmore. High-throughput powder diffraction. iv. cluster validation using silhouettes and fuzzy clustering. *Journal of Applied Crystallography*, 37(6):874–882, 2004.
- [10] G. Barr, W. Dong, C. J. Gilmore, A. Parkin, and C. C. Wilson. *dsn timer*: A computer program to cluster and classify cambridge structural database searches. *Journal of Applied Crystallography*, 38(5):833–841, 2005.
- [11] I. J. Bruno, J. C. Cole, P. R. Edgington, M. Kessler, C. F. Macrae, P. McCabe, J. Pearson, and R. Taylor. New software for searching the cambridge structural database and visualizing crystal structures. *Acta Crystallographica Section B*, B58(3 Part 1):389–397, 2002.
- [12] C. Burt. *The Factors of the Mind*. London University Press, 1940.
- [13] A. Collins, G. Barr, W. Dong, C. J. Gilmore, D. S. Middlemiss, A. Parkin, and C. C. Wilson. The application of cluster analysis to identify conformational preferences in enones and enamines from crystal structural data. *Acta Crystallographica Section B*, 63(3):469–476, 2007.
- [14] A. Collins, A. Parkin, G. Barr, W. Dong, C. J. Gilmore, and C. C. Wilson. Identifying structural motifs in intermolecular contacts using cluster analysis part 2. interactions of carboxylic acids with secondary amides. *CrystEngComm*, –:–, 2007.
- [15] A. Collins, A. Parkin, G. Barr, W. Dong, C. J. Gilmore, and C. C. Wilson. Configurational and conformational classification of pyranose sugars. *Acta Crystallographica Section B*, 64(1):57–65, 2008.

- [16] A. L. Comrey. *A first Course in Factor Analysis*. Academic Press, 1973.
- [17] H. G. S. ConsortiumInternational. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–945, Oct. 2004.
- [18] T. F. Cox and M. Cox. *Multidimensional Scaling, Second Edition*. Chapman & Hall/CRC, 2 edition, 2000.
- [19] H. S. M. Coxeter. *Introduction to Geometry*. Wiley Classics Library, 1989.
- [20] A. dAspremont, L. El Ghaoui, M. I. Jordan, and G. R. G. Lanckriet. Dspca: Sparse pca using semidefinite programming. <http://www.princeton.edu/~aspremon/DSPCA.htm>, Last visited: 30 September 2009.
- [21] A. d’Aspremont, L. E. Ghaoui, M. I. Jordan, and G. R. G. Lanckriet. A direct formulation for sparse pca using semidefinite programming. *SIAM Review*, 49(3):434 – 448, 2007.
- [22] C. S. Database. Csd entries: Summary statistics. http://www.ccdc.cam.ac.uk/products/csd/statistics/entry_stats.php4, page Last visited : September 2009, January 2009.
- [23] B. S. Everitt, S. Landau, and M. Leese. *Cluster Analysis, Fourth edition*. A Hodder Arnold Publication, 2001.
- [24] H. D. Flack and G. Bernardinelli. Absolute structure and absolute configuration. *Acta Crystallographica Section A*, 55(5):908–915, Sep 1999.
- [25] K. R. Gabriel. The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, 58(3):453–467, 1971.
- [26] C. J. Gilmore, G. Barr, and J. Paisley. High-throughput powder diffraction. i. a new approach to qualitative and quantitative powder diffraction pattern analysis using full pattern profiles. *Journal of Applied Crystallography*, 37(2):231–242, 2004.

- [27] J. Gower and G. Dijksterhuis. *Procrustes Problems*. Oxford University Press, 2004.
- [28] J. Gower and D. Hand. *Biplots*. Chapman and Hall, 1996.
- [29] J. C. Gower. Three-dimensional biplots. *Biometrika*, 77(4):773–785, 1990.
- [30] C. Hansch, S. H. Unger, and A. B. Forsythe. Strategy in drug design. cluster analysis as an aid in the selection of substituents. *Journal of Medicinal Chemistry*, 16(11):1217–1222, Nov. 1973.
- [31] H. Harman. *Modern Factor Analysis*. Chicago University Press, 3rev ed edition, 1976.
- [32] S. Henikoff and J. G. Henikoff. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America*, 89(22):10915–10919, 1992.
- [33] D. G. Higgins and P. M. Sharp. Clustal: a package for performing multiple sequence alignment on a microcomputer. *Gene*, 73(1):237 – 244, 1988.
- [34] R. Hurley, J.R.and Cattell. Producing direct rotation to test a hypothesized factor structure,. *Behavioral Science*, 7:258262, 1962.
- [35] R. Jarvis and E. Patrick. Clustering using a similarity measure based on shared near neighbors. *IEEE Transactions on Computers*, 22:1025–1034, 1973.
- [36] I. Jolliffe. *Principal Component Analysis*. Springer Series in Statistics, 2nd edition edition, 2002.
- [37] H. F. Kaiser. The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23:187–200, 1958.
- [38] M. Kessler, J. Pérez, M. C. Bueso, L. García, E. Pérez, J. L. Serrano, and R. Carrascosa. Probabilistic model-based methodology for the conformational study of cyclic systems: application to copper complexes

- double-bridged by phosphate and related ligands. *Acta Crystallographica Section B*, 63(6):869–878, Dec 2007.
- [39] M. Larkin, G. Blackshields, N. Brown, R. Chenna, P. McGettigan, H. McWilliam, F. Valentin, I. Wallace, A. Wilm, R. Lopez, J. Thompson, T. Gibson, and D. Higgins. Clustal w and clustal x version 2.0. *Bioinformatics*, 23(21):2947–2948, 2007.
- [40] C. F. Macrae, I. J. Bruno, J. A. Chisholm, P. R. Edgington, P. McCabe, E. Pidcock, L. Rodriguez-Monge, R. Taylor, J. van de Streek, and P. A. Wood. *Mercury CSD 2.0* – new features for the visualization and investigation of crystal structures. *Journal of Applied Crystallography*, 41(2):466–470, Apr 2008.
- [41] E. R. Malinowski. *Factor Analysis in Chemistry, 2nd Edition*. Wiley-Interscience, 2 edition, 1991.
- [42] B. F. Manly. *Multivariate Statistical Methods: A primer*. Chapman & Hall/CRC, 3rd edition, 2005.
- [43] Mathworks. Adapted from:<http://www.mathworks.com/products/demos/statistics/factorandemo.html>. *Last Visited: 30 September 2009*.
- [44] R. Mojena. Hierarchical grouping methods and stopping rules: an evaluation. *The Computer Journal*, 20(4):359–363, 1977.
- [45] P. Murray-Rust and J. Raftery. Computer analysis of molecular geometry : Part vi: Classification of differences in conformation. *Journal of Molecular Graphics*, 3:50–59, 1985.
- [46] P. Murray-Rust and J. Raftery. Computer analysis of molecular geometry, part vii: the identification of chemical fragments in the cambridge structural data file. *Journal of Molecular Graphics*, 3:60–68, 1985.
- [47] A. Parkin, G. Barr, W. Dong, C. J. Gilmore, and C. C. Wilson. Identifying structural motifs in inter-molecular contacts using clusteranalysis part 1. interactions of carboxylic acids with primary amides and with othercarboxylic acid groups. *CrystEngComm*, 2006,, 8:257–264, 2006.

- [48] C. Pascard. Small-molecule crystal structures as a structural basis for drug design. *Acta Crystallographica Section D*, 51(4):407–417, Jul 1995.
- [49] K. Pearson. Contributions to the mathematical theory of evolution. ii. skew variation in homogeneous material. *Philosophical Transactions of the Royal Society of London. (A.)*, 186:343–414, 1895.
- [50] J. Pérez, K. Nolsøe, M. Kessler, L. García, E. Pérez, and J. L. Serano. Bayesian methods for the conformational classification of eight-membered rings. *Acta Crystallographica Section B*, 61(5):585–594, 2005.
- [51] P. J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53 – 65, 1987.
- [52] C. Spearman. "general intelligence," objectively determined and measured. *The American Journal of Psychology*, 15(2):201–292, 1904.
- [53] C. Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904.
- [54] SPSS. Spss for windows. rel 15.0.0. <http://www.spss.com/uk/>, Last Visited: September 2009.
- [55] G. Thomson. *The Factorial Analysis of Human Ability*. London University Press, 1951.
- [56] L. Thurstone. *Multiple Factor Analysis*. University of Chicago, 1947.
- [57] M. Vingron and P. R. Sibbald. Weighting in sequence space: a comparison of methods in terms of generalized sequences. *Proceedings of the National Academy of Sciences of the United States of America*, 90(19):8777–8781, 1993.
- [58] E. W. Weisstein. "heron's formula." from mathworld—a wolfram web resource. <http://mathworld.wolfram.com/HeronsFormula.html>, Last accessed 29th September 2009.

- [59] Z. F. Weng, W. D. S. Motherwell, F. H. Allen, and J. M. Cole. Conformational variability of molecules in different crystal environments: a database study. *Acta Crystallographica Section B*, 64(3):348–362, 2008.