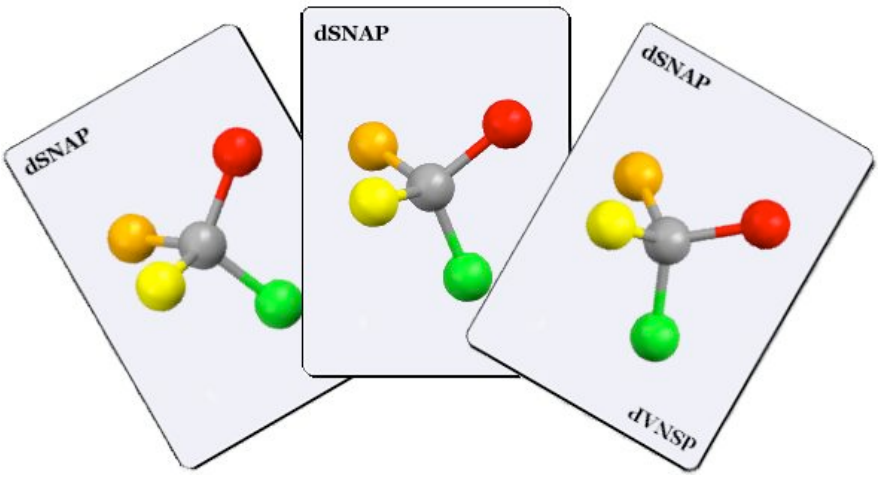
*d*SNAP

A computer program to cluster and classify results from Cambridge Structural Database searches



Tutorial

University of Glasgow

Version 1.0.0
April 2009

About *d*SNAP

*d*SNAP is a program for comparing the geometries of groups of atoms extracted from crystal structures in the Cambridge Structural Database. Because the database contains so many structures, if can be difficult and time-consuming to extract meaningful structural information, particularly for large datasets; *d*SNAP can facilitate this process.

*d*SNAP can be used to investigate intramolecular geometry, or to analyse intermolecular interactions when more than one chemical residue is defined in a single database search. The method is suitable for both organic and organometallic systems. The aim is to distinguish different geometries through the use of cluster analysis.

Analysis can be performed on groups of 3-20 atoms, for datasets of up to 4000 instances of the search fragment. The user can also choose to include their own structures from cifs in the analysis to see how they compare to structures already in the database.

As *d*SNAP knows no chemistry, it minimises the risk of introducing bias into the analysis of results.

Introduction to the Tutorial

This tutorial consists of three worked examples which aim to introduce some of the main features of *d*SNAP.

Examples 1 and 2 involve an organic and organometallic example respectively. Both examples demonstrate how you can include further structures from your own cifs. It is recommended that you work through only one of these systems, so pick the one that you feel is the more relevant to your own research. The discussion will focus on setting up the database search, input of the data into *d*SNAP and explanation of some of the tools available for analysing the results.

Example 3 demonstrates a system involving intermolecular interactions. Although this system involves only organic materials, the principles are also applicable to organometallic systems. The discussion material will focus on setting up the ConQuest search and analysis of the clustering results.

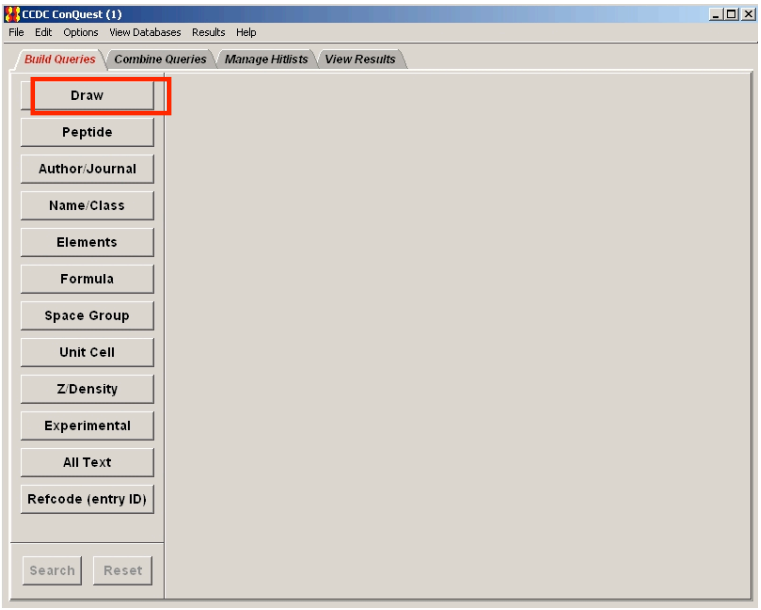
Each example has detailed instructions, along with some tips and notes.

All of the examples for this tutorial consist of quite small data sets, so that results can be generated quickly.

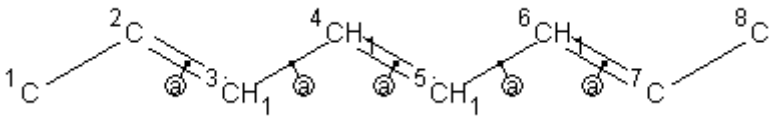
Example 1:
A simple organic fragment

Step 1: Set up the CSD search using *ConQuest*

Open *ConQuest* and click on the **Draw** button in the **Build Queries** tab.



Draw the fragment shown below:



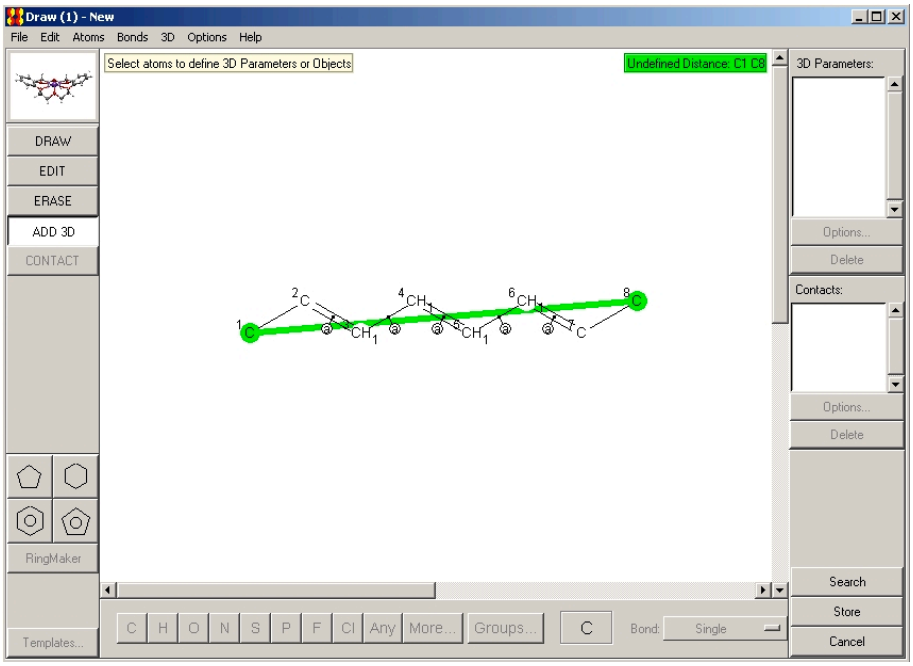
Before starting the search, define at least one interatomic distance, angle or torsion angle by clicking on the **ADD 3D** button on the left side of the screen and then selecting the atoms involved (*e.g.* C1-C8 as DIST1). Click ***Define***, then ***Done***.

This must be done so that that the required files can be generated from the search results. However, it does not matter which parameter is chosen or which atoms it involves.

Tips:
The order in which the atoms are drawn corresponds to how they are numbered, and this numbering scheme is later used in *d*SNAP. For ease in following this tutorial, it is recommended that you draw the fragment to get the same numbering scheme.

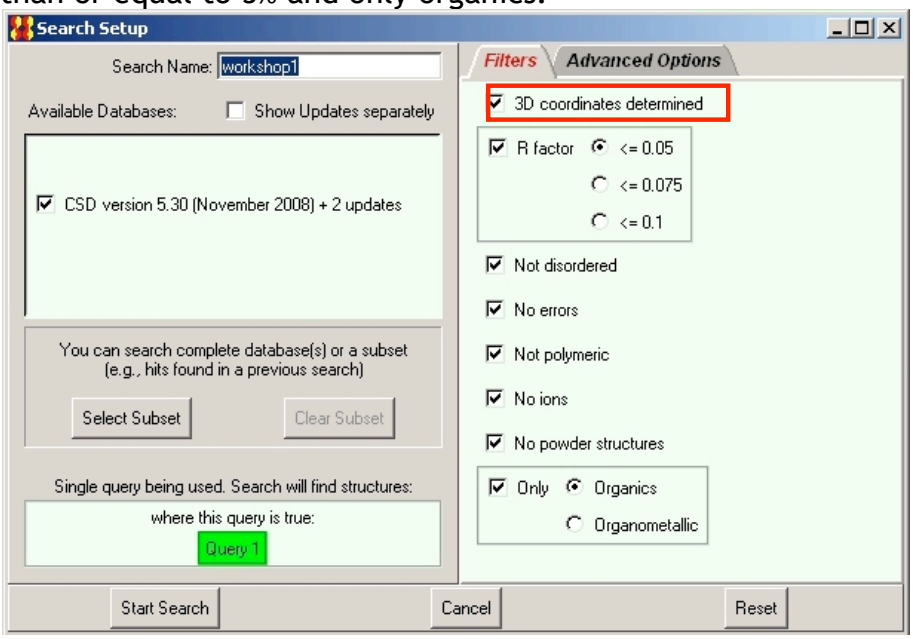
The circled 'a' label attached to the double bonds indicates that they have been defined to be acyclic. This can be done by right-clicking on a bond and selecting **Cyclicity > Acyclic** from the drop-down menu. Alternatively, from the toolbar menu choose **Bonds > Cyclicity > Acyclic** and click on each of the bonds that is required to be acyclic in turn. These are listed in the window. Then click *OK*.

The hydrogen atoms have been defined implicitly, that is no bonds are drawn in connecting them to the carbon atoms. To do this, right-click on the atom to which hydrogens are to be added, selected **Hydrogens > Generate**. This will generate the number of hydrogens permitted by the valence of the atom. The number of hydrogens can also be specified. Alternatively from the toolbar menu choose **Atoms > Hydrogens > Generate** and click on each of the atoms that that you wish to have hydrogens on in turn. These are listed in the window. Then click *OK*.



Tip:
If there is a parameter that you are particularly interested in, if you define that you will then have the information readily available for exporting to other programs, *e.g.* Vista or Excel.

Click the **Search** button. Change the name of the search to *workshop1*. Always make sure the ***3D coordinates determined*** box is checked in the search dialog box. For this example, also check all the other boxes, with the R-factor less than or equal to 5% and only organics.



Note:
The maximum number of fragments that can be analysed by *d*SNAP is 4000. Note that the number of hit structures is generally smaller than the number of hit fragments; some structures contain multiple instances of the hit fragment. It may be necessary to introduce further restrictions on the database search if the data set is large. This can be done in a number of ways, *e.g.* adding more atoms to the search fragment, including hydrogen atoms, placing stricter limits on the value of the R-factor, on experimental data collection temperature, or when the crystal structure was published.

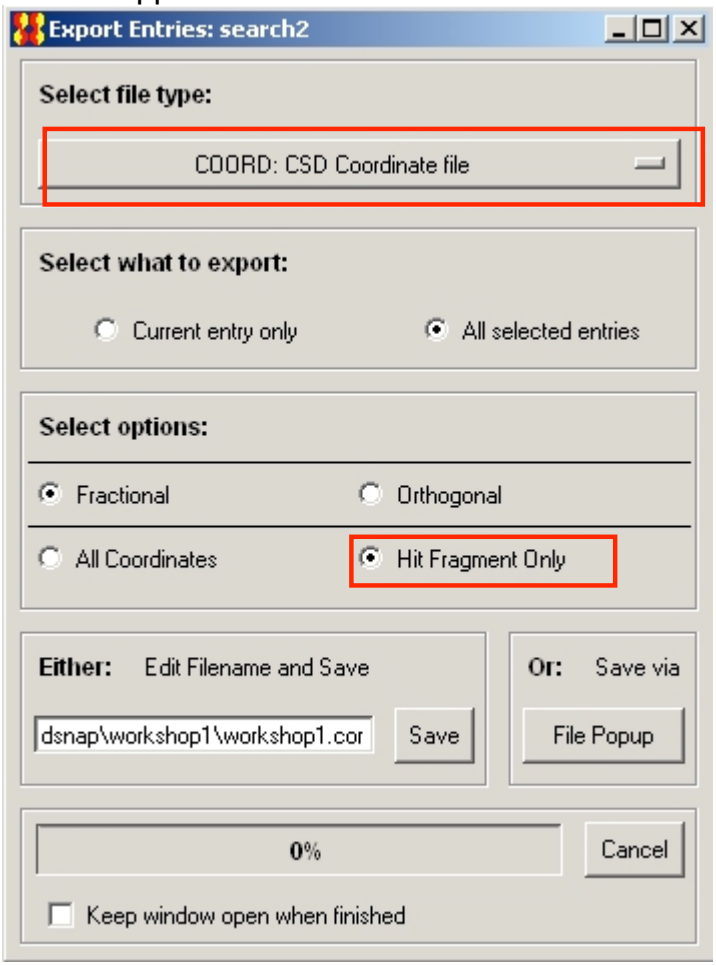
Start the search.

Step 2: Save the required files

Two files are needed for input into *d*SNAP: a .cor file, which contains a list of the coordinates of the atoms in the fragment, and a .fgd file, which contains the information about atom types, how the atoms are connected to each other, and other

atom and bond definition information (*e.g.* whether a bond has been defined as cyclic or acyclic, or if the connectivity of an atom has been constrained). This file also contains the atom numbering scheme for the fragment.

First, create the .cor file. Go to **File > Export Entries as…** A dialog box will appear.



Select file type **COORD: CSD Coordinate file**.

Make sure that you check the box for ***Hit fragment only*** in the **Select options** section, as *d*SNAP will not run if the coordinates of all the atoms in the asymmetric unit are exported (an error message to this effect will be displayed in *d*SNAP if this is done by accident).

Save this into the *workshop1* folder.

Now create the fgd file. Go to **File > Export Parameters and Data**. Use the default options.

The fgd file is created along with 2 other files, .fgn and .tab, but these are not required and can be ignored or deleted.

Notes:
This folder already contains a cif, which in this example will also be imported for inclusion in the cluster analysis.

If you are not planning on comparing your own structures to database structures, you do not need a cif.

Make sure that all the files are saved with the same root filename (in this case, workshop1).

Keep *ConQuest* open until the *d*SNAP cluster analysis has run successfully.

Tip:
It is strongly recommended that you save the search (as a .cqs file) for future reference.

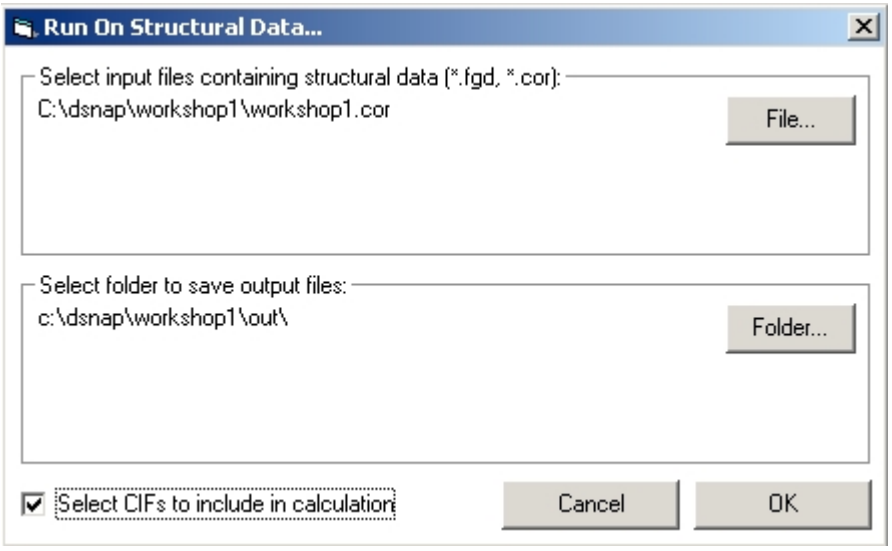
Step 3: Open *d*SNAP and identify the input files

Double-click on the *d*SNAP icon to launch the program.

When the dialog window appears, click on the **Run new analysis** button. A new window will open. Click the **File…** button to browse for the correct folder containing the .cor and .fgd files and select one of them (it does not matter which).

Create the output folder. Note that the output cannot be saved in the same folder as where the input files are saved. It is recommended that you create a subfolder in the input directory for the output. Click the **Folder…** button and browse for the input folder. To select a folder, double-click on it. For the current working directory, the icon will appear as an open folder.

Check the box at the bottom left to include a cif.



Verify that the path for the output files is correct, then click ***OK***. A new dialog box will appear asking for the name of the cif. Select *bca1.cif* from the *workshop1* folder and click ***OK***.

Another window will appear. This is used to tell the program which atoms in the cif correspond to which atoms from the original search fragment.

Note:
If you wanted to include more than one cif, this can be done using standard multiple selection options in the **Open CIF** dialog box.

The name of the cif will be truncated to the last 7 characters of the root filename to create its identifier in dSNAP. It is recommended that when you use your own cifs, you save a copy of each in the same folder as your fgd and cor input files, giving it a meaningful 7-character filename. This is particularly useful if you wish to import more than one cif, as all cifs must be located in the same folder.

To associate an atom in the cif with the corresponding atom in the search fragment, right-click on the appropriate atom in the cif atom list and select the search fragment atom from the drop-down list that appears. As each atom is assigned, the name of the cif atom is written (in italics) below the atom name in the diagram.

To work out which atoms are which, open the cif in *Mercury* and view the atom labels. In this case there is only one instance of the search fragment in the asymmetric unit.

For this cif, the atom assignments are given in the table below:

CIF atom	Search fragment atom
C2	C1
C3	C2
C4	C3
C5	C4
C6	C5
C7	C6
C8	C7
C9	C8

Note:
The cif used in this example is taken directly from the CSD. Here, this helps to provide a check that all the atoms have been correctly assigned.

As each atom is assigned, the table on the right of the screen is also updated with which search fragment atom corresponds to which cif atom. When all the atoms have been assigned, check that the table is correct, then click ***OK***. You will be asked if you want to include another fragment from this cif. Click ***No***.

This fragment has topological symmetry, *i.e.* there are some atoms that are equivalent to each other on the basis of the chemical connectivity of the fragment. In this case it affects all the atoms in the fragment, but for other choices of search fragments, only some of the atoms in the fragment may be related by local symmetry. The presence of topological symmetry in a fragment causes an ambiguity in the way in which the atoms in the fragment which is extracted from the CSD are numbered. If this ambiguity is not dealt with, it can result in the presence of false clusters in the cluster analysis and so any local symmetry in the fragment must be accounted for. This is done automatically by *d*SNAP for all fragments, including fragments taken from cifs.

The output for this topological symmetry correction is contained in the *pre_processing_log.txt* file, which is written to the output folder. This gives a list of the possible ways in which the atom numbering can be permuted, which are known as symmetry possibilities.

Note:
There is a maximum limit of 4000 symmetry possibilities. In practical terms, this means that 7-fold or higher order symmetry cannot be run with an automatic symmetry correction.

For example, this fragment has two symmetry possibilities, given by:

Symm Poss	1	2	3	4	5	6	7	8
1	1	2	3	4	5	6	7	8
2	8	7	6	5	4	3	2	1

The output also lists which symmetry possibility has been assigned to each fragment. Looking at the first ten hit fragments:

Refcode	Symm Poss output
BASBIH	1
BASCAA	2
CROCAL_01	1
CROCAL_02	1
CROCAL_03	1
DIXJIE	2
DIYWUD	1
DPHOCE03	1
FITFIY_01	1
FITFIY_02	1

Here, those fragments with the symmetry possibility output of 2 have been renumbered, so in those fragments C1 has become C8, C2 is C7, C3 is C6, etc., while fragments with a symmetry possibility of 1 have not had their atom numbering altered.

A dialog box will appear asking if you wish to correct for topological symmetry. Click ***OK***. The cluster analysis will start.

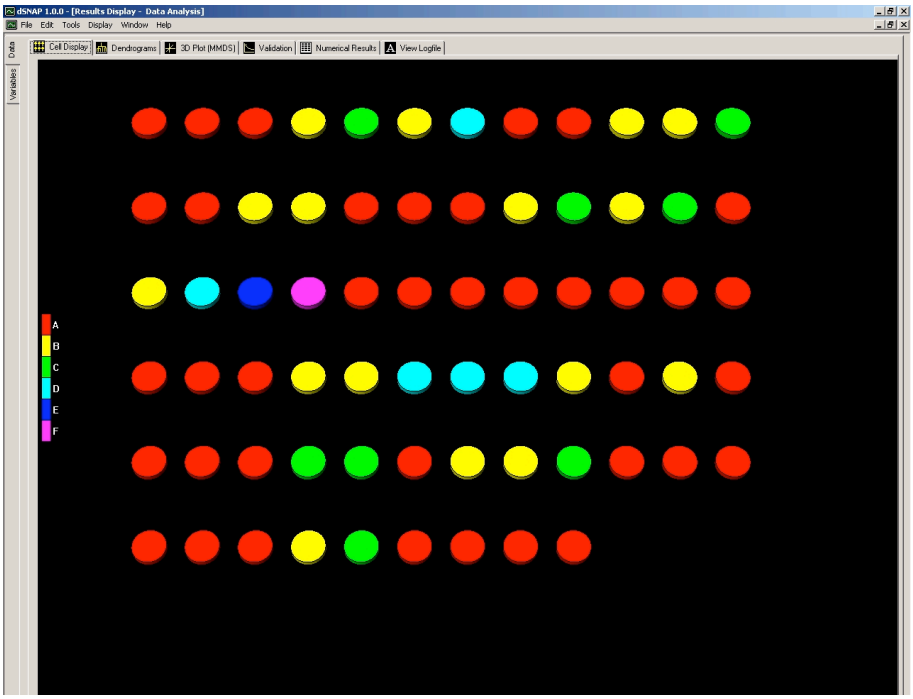
Tip:
This example should run quite quickly, but some analyses, particularly those involving a large number of hits from the CSD, a large number of atoms in the search fragment, or high orders of topological symmetry, can take several hours to run, and may be best run overnight.

Step 4: Viewing and analysing the results

Progress bars appear as *d*SNAP carries out different stages of the clustering process.

The results are displayed in a selection of analysis tools, which can be accessed by a series of tabs at the top of the screen. Results can be displayed in **Data** space, which relates to the correlations between fragments, or **Variables** space, which relates to the correlations between parameters. The vertical tabs on the left of the screen are used to switch between them.

The first screen to appear will be the **Cell Display** in **Data** space, which will look something like this:



Each circle represents a hit fragment. Those which are the same colour are currently assigned to the same cluster. The key down the left side of the screen associates each cluster colour with a letter (for up to 30 clusters; beyond this point it can be hard to distinguish the different colours).

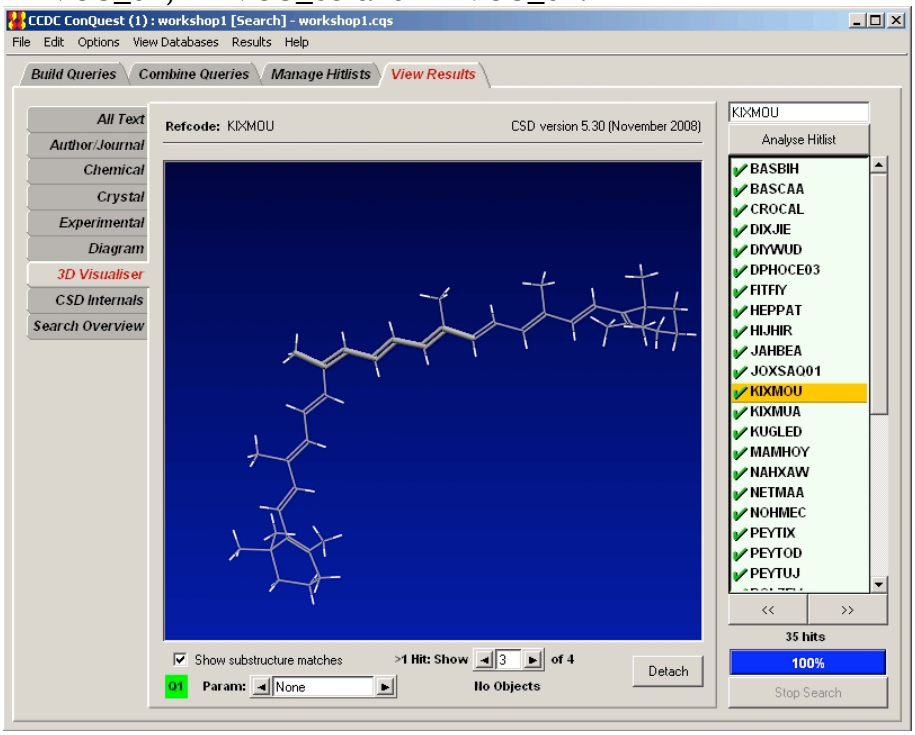
Each fragment has a label. By default these are not shown for datasets of more than 15 fragments. However, by right-clicking anywhere in the display, the labels can be shown for all fragments (***Show All Labels***) or for selected fragments (***Show Selected Labels***). Alternatively, when the mouse is hovering over a fragment, you will see a tooltip box with the fragment name.

Note:
These options for viewing labels are also available in the other graphics tabs.

The fragments are displayed in the Cell Display in the order in which they are output from the CSD. This is generally alphabetically, but structures which have been included from updates to the database are found at the end. Fragments from included cif files appear last. These are also easy to spot as their identifying code will begin with a '+'.

The number of hit fragments is bigger than the number of hit refcodes from the CSD search. This is because the search fragment occurs more than once in some of the structures, due to multiple instances of the same fragment in a single molecule, or because of multiple molecules, $Z'>1$, in the asymmetric unit of the structure. When this occurs, the refcode has *nn* appended to it, where *nn* is the number of the

fragment in that structure. If you refer back to the original CSD search in ConQuest you can highlight these hits individually, *e.g.* refcode KIXMOU, which has 4 sub-structure matches for the search fragment, labelled KIXMOU_01, KIXMOU_02, KIXMOU_03 and KIXMOU_04.



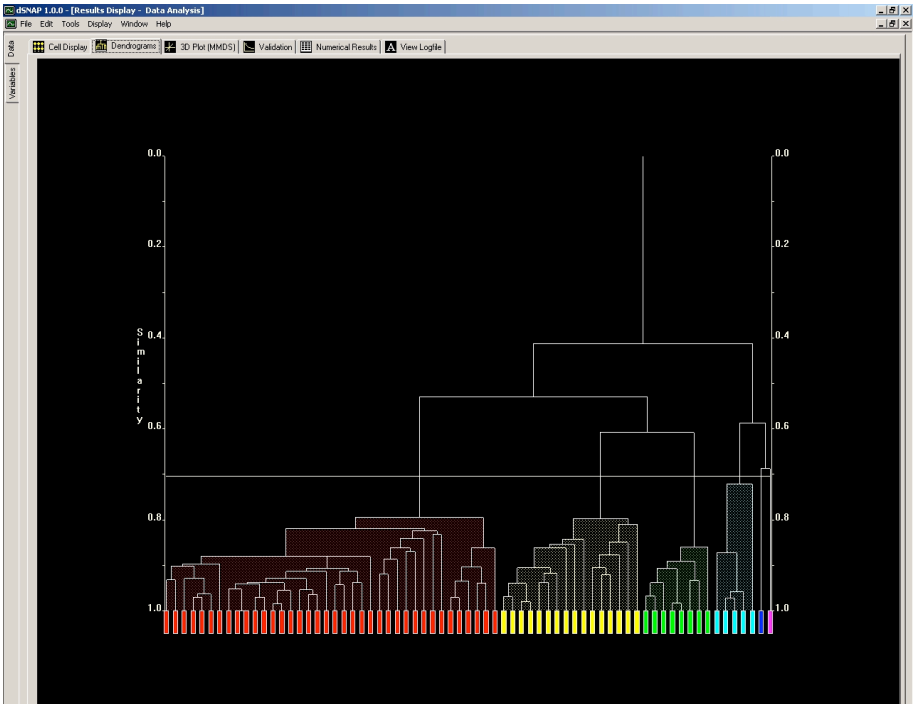
Tip:
Here, the third hit fragment is being highlighted as capped sticks, while the rest of the structure is shown as wireframe in the *ConQuest* **3D Visualiser**. In *d*SNAP, this hit will correspond to KIXMOU_03. This display option is accessed by right-clicking in the *ConQuest* display window and picking **Highlight Hit > Hit as Capped Stick.**

Fragments in the Cell Display in *d*SNAP can be selected by clicking on them. Multiple fragments can also be selected. If the fragments are in sequence, click on the first, hold down the *Shift* key and click on the last. If the fragments to be selected are not in sequence, hold down the *Ctrl* key and click on each of the fragments of interest. An entire cluster can be selected by clicking on the coloured box corresponding to the cluster in the key on the left of the screen. Multiple clusters can be selected by clicking on the each cluster while holding down either the *Shift* or the *Ctrl* key. Note each cluster must be selected individually.

The selected hits can be viewed in Mercury, either by pressing F3, or through the menu. *Go to* ***Tools > Show Selected Hits in Database Viewer… F3***.

In *d*SNAP, how the clusters are assigned can be seen and adjusted by viewing the dendrogram.

Click on the **Dendrogram** tab. The dendrogram being displayed should look something like this:



Each box at the bottom of the screen represents a single hit fragment. The boxes are joined by horizontal lines called tie-bars, which link together fragments according to the calculated similarity between each connected branch. The vertical axis is a similarity scale, with zero similarity at the top, and a similarity of 1.0 at the bottom ie if two fragments are joined by a tie-bar near the bottom of the dendrogram then they can be considered very similar, justifying their being grouped together. If two branches do not meet until near the top of the dendrogram, the associated fragments are only loosely related to each other.

The horizontal line that spans the width of the dendrogram marks the cut level. Any fragments that are linked at higher level of similarity than the cut level (*i.e.* lower down the dendrogram) are put into the same cluster), while those which are less similar (*i.e.* their tie-bars lie above the cut level on the dendrogram) belong to separate clusters. The initial cut level is determined using Principal Components Analysis, but can be adjusted by left-clicking anywhere in the dendrogram display and scrolling using the scroll wheel on the mouse.

Moving the line up (raising the cut level) will decrease the number of clusters, and overall the fragments within each cluster will be less similar to each other, while lowering it will increase the number of clusters, and the fragments in each cluster will be more similar to one other.

See the effect of raising and lowering the cut level on the numbers of clusters. Then click on the **3D Plot (MMDS)** tab.

Notes:
For datasets with a very large number of hit fragments the individual boxes may not be distinguishable and you may need to zoom in to see the finer details.

Because the imported cif was taken from a CSD structure which is also included in this analysis, it can be identified as it is linked to another fragment with a similarity of 1.

In cases where the cif atoms have been assigned incorrectly, the fragment will frequently appear in a cluster by itself.

Note:
Alternatively, if you do not have a scroll wheel on the mouse (*e.g.* using a laptop) the cut level can be adjusted by holding down the *Alt* key, left click on the cut level and drag the cut level up or down). It can also be adjusted via the menu: **Tools > Set Cut-Level To…** and then enter the required value. This also allows the cut-level to be set to a specific value.

At this stage, do not save the changes to the dendrogram when prompted.

The plot should look something like this:

Each sphere represents a hit fragment. Fragments that are located close together in space can be expected to have similar geometries. The colours are taken from the dendrogram. Ideally, spheres of the same colour will be close together and well separated in space from spheres of different colours. This indicates good agreement between the MMDS plot and the dendrogram. The plot can be rotated in space by dragging while holding down the left mouse button.

Some of the spheres have white spikes coming from them. These indicate the most representative fragment in each cluster; this can be useful for selecting examples to illustrate the geometries of each cluster. They are only shown for clusters containing three or more fragments.

Note:
The most representative sample highlighting can be switched off by right-clicking anywhere within the 3D Plot display and clicking on ***Show MRP Marks***.

Under the **Validation** tab are several tools for analysing results. These are not so relevant to this tutorial, but are explained in the *d*SNAP manual.

In order to assess whether the default cut level is the most appropriate, first compare the geometries of the fragments within each cluster to see whether they are similar enough* to be included in the same cluster. Then compare the clusters to one another.

*** 'Similar enough' is subjective, and very much dependent on the level of detail that is relevant and appropriate to the investigation. At the two extremes, all fragments could be considered in a single cluster as they all have the same connectivity, or all fragments whose similarity is less than 1 are different as they do have different geometries, no matter how small the differences are. For most datasets, a situation somewhere between these is probably the most appropriate or meaningful, and it is the justification of cut level that is important. The initial cut level provides a good starting point for the number of clusters in the data set, but it may need to be adjusted depending on the level of detail that is required from the analysis.**

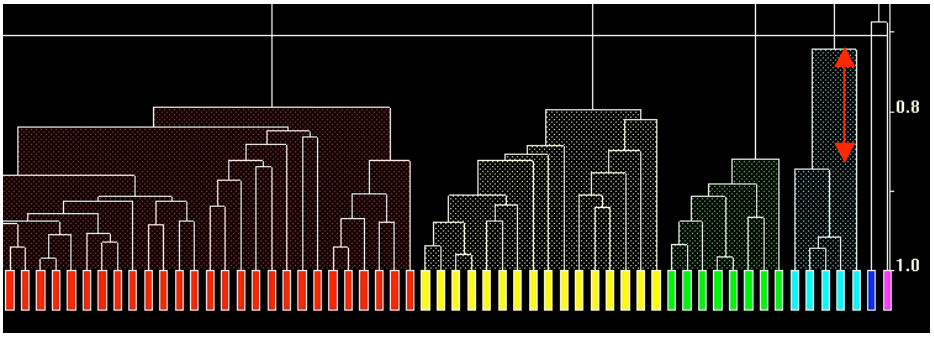
One of the things to note is that in the MMDS plot all of the groups of the same coloured spheres are very tightly packed except the cyan cluster. Switch to the Dendrogram tab. Highlight all the fragments in the cyan cluster and press ***F1***. Dismiss the window that appears displaying the 2D chemical diagram of the fragment by clicking *OK*. This brings up the **Multiple Fragments Viewer**, which displays the optimum overlay of all the selected fragments.

The fragment viewer is a very useful tool in establishing the suitability of a chosen cut level as it provides an excellent check of how similar the fragments are when viewed en masse. Small differences that may be lost when comparing the structures individually can become much more apparent. This allows outliers for a given cluster, or the dataset as a whole, to be identified much more rapidly.

Although the fragments appear to be overlaying well, there is one fragment that has a slightly different geometry.

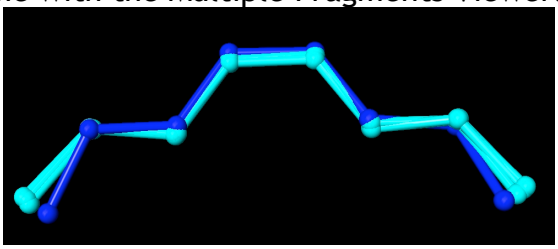
Close the fragment viewer. Looking at the dendrogram, there is quite a big step between the topmost tie-bar in the cyan cluster and the next highest tie-bars. This is a good indication that it may be appropriate to lower the cut level and split the cluster further into two groups.

Tips:
An easy way to do this is to switch to the Cell Display, left-click on ***D*** in the key. This selects the whole cluster. Then switch to the dendrogram display. They do not appear selected, but pressing *F1* will bring up the fragment viewer for the whole cluster.
The Multiple Fragment Viewer can also be accessed through the menu. Go to **Tools > Show Selected Fragments in 3D Viewer… F1**.



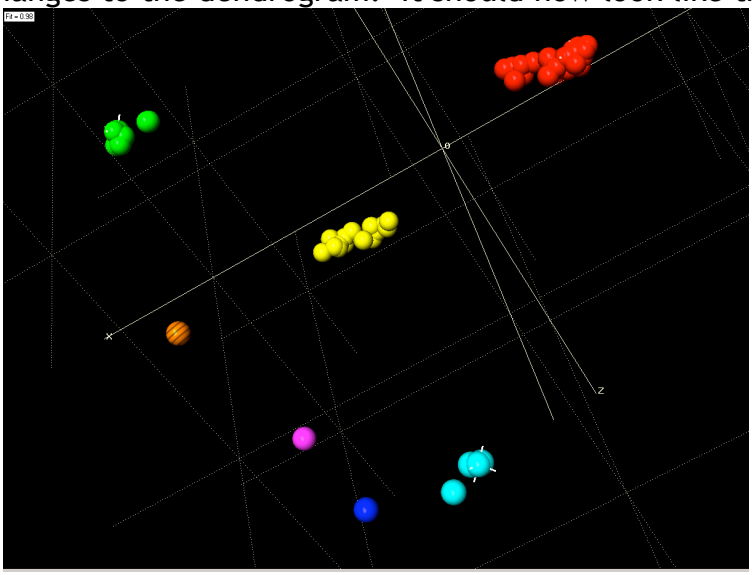
Now, change the cut level in the dendrogram from 0.704 to around 0.764, the cyan cluster becomes separated into two groups, coloured cyan and blue. Notice that the cluster that was previously blue is now coloured pink, and the cluster that was pink is now brown striped.

View all the fragments in the current cyan and blue clusters at the same time with the Multiple Fragments Viewer.



The fragment in the blue cluster has a geometry that is distinct from the fragments in the cyan cluster.

Close the fragment viewer and click on the 3D Plot tab, saving the changes to the dendrogram. It should now look like this:

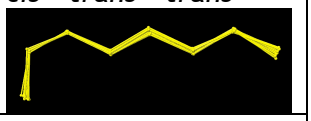
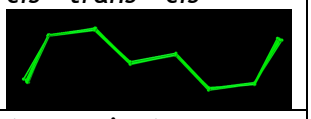
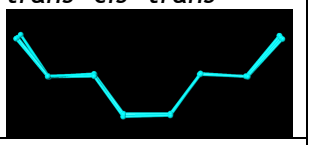
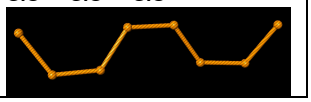


Go back to the Dendrogram tab and view each cluster in turn using the Multiple Fragments Viewer to verify that the groupings make sense.

Note:
This image shows a zoomed-in and rotated view of the MMDS plot, with the sphere size decreased.

Sphere size can be adjusted by holding down the left mouse button and dragging up or down on the screen while pressing *Alt*.

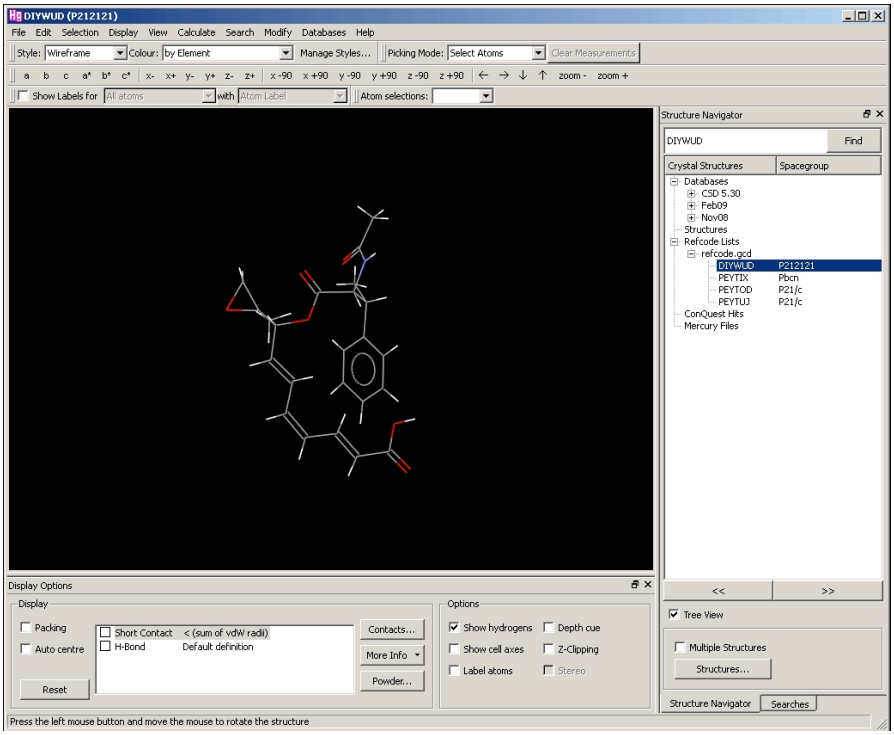
To zoom in on a display, drag a box round the area to be enlarged while holding down the left mouse button and the *Shift* key.

Cluster	No. of fragments	Description of fragment geometry
A (red)	38 (inc. *bca1.cif*)	*trans - trans -trans* 
B (yellow)	16	*cis - trans - trans* 
C (green)	8	*cis - trans - cis* 
D (cyan)	4	*trans -cis -trans* 
E (blue	1	*trans - cis - trans* 
F (pink)	1	*cis - cis - trans* 
G (brown stripe)	1	*cis - cis - cis* 

It is now necessary to verify that the clusters make chemical sense. This can be done by viewing not just the fragment but the whole molecule, or if necessary, the entire unit cell. This can easily be done using Mercury.

In this data set, we may be particularly interested in the differences between the cyan and blue clusters. In the Dendrogram tab, highlight the structures in the cyan cluster and press ***F3***. By default this opens *Mercury*.

Note:
The numbers of fragments in a cluster at a given saved cut level are given in the logfile, ***SNAPlog.txt***. This can also accessed via the **View Logfile** tab.



Note:
The option to view the complete structure is also available through the menu. Go to **Tools > Show Selected Hits in Database Viewer… F3**. This option can be also be accessed from the Cell Display and 3D plot in Data space, and when a fragment is selected in a scatterplot generated in the Numerical Results tab in Variables space.

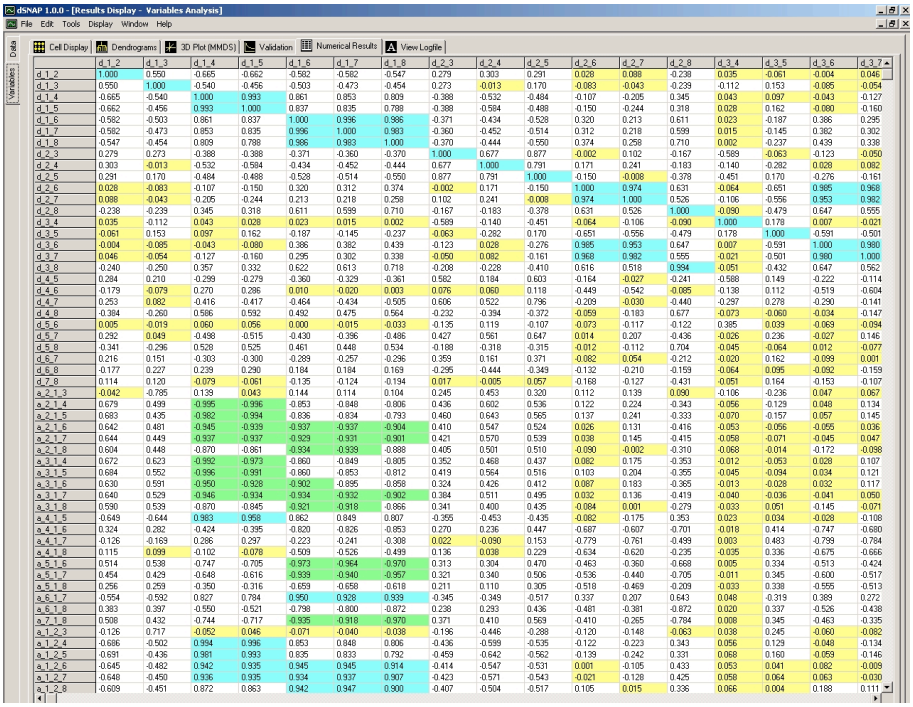
All the refcodes for the structures in the cluster are listed. Look at each in turn. Then look at the fragment in the blue cluster.

What we notice is that in the blue cluster the centre of the C4-C5 bond lies on a 2-fold rotation axis of the space group and the molecule possesses C_2 point symmetry, whereas the fragments in the cyan cluster are found in molecules that do not possess internal symmetry overall.

Go to the **Numerical Results** tab and then click on the **Variables** tab on the left of the screen.

This brings up a matrix giving the correlations between all possible pairs of parameters (the distances and angles), ranging between –1 to +1. The diagonal is always +1. The matrix can be colour-coded for different values of the correlation, using up to 5 colours.

Note:
Viewing the Numerical Results in Data space displays the correlation matrix between each fragment.



Clicking on the correlation for a pair of parameters brings up a scatter plot. In the plot, each dot represents a fragment. They are colour-coded according to cluster, as determined by the choice of cut level in the last saved dendrogram. Hovering over the fragment brings up a tool-tip box with the name of the fragment and the value of each of the two parameters for that fragment. Clicking on a dot selects it and opens *Mercury* to display it.

We will use the Numerical Results to compare the bonded distances and the bond angles. Clicking on a cell brings up a scatterplot relating the two parameters.

Notes:
The choice of colours and the range for each colour can be changed in the **Edit > Options** menu. Click on the ***Select Correlation Matrix Colours*** box at the bottom left of the window. This brings up a new window headed ***Numerical Results Formatting***. Left-clicking on the colour boxes brings up a colour selection box. The use of dark colours is not recommended as the values of the correlations will become difficult to read. The colours will be up-dated the next time that the Numerical Results are viewed in Variables space after switching from Data space.

Including the colours means this tab takes longer to load, and so it is switched off by default if there are more than 500 parameters (this corresponds to a 10 atom fragment). You can increase this value if you wish to colour-code the correlation matrix for larger fragments. This table gives the number of parameters for different numbers of atoms in the fragment, up to the program's maximum limit of 20 atoms.

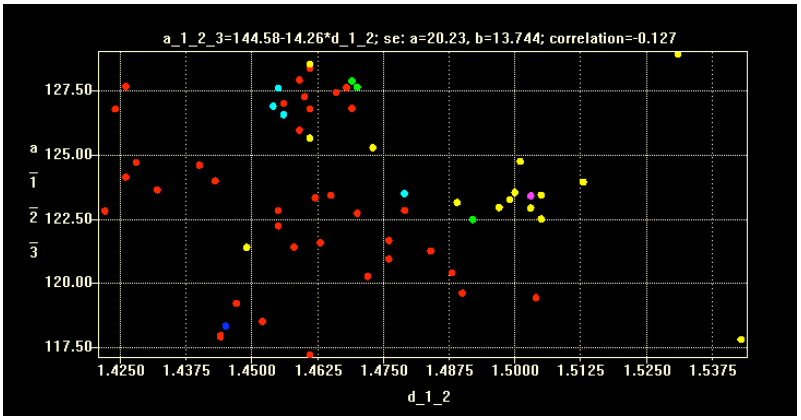
Number of atoms	Number of parameters
11	550
12	726
13	936
14	1183
15	1470
16	1800
17	2176
18	2601
19	3078
20	3610

So if we wish to view the relationship between the C1-C8 distance and the C1–C2 distance, we could click in either of the cells highlighted below:

	d_1_2	d_1_3	d_1_4	d_1_5	d_1_6	d_1_7	d_1_8	d_2_3	d_2_
d_1_2	1.000	0.550	-0.665	-0.662	-0.582	-0.582	-0.547	0.279	0.303
d_1_3	0.550	1.000	-0.540	-0.456	-0.503	-0.473	-0.454	0.273	-0.01
d_1_4	-0.665	-0.540	1.000	0.993	0.861	0.853	0.809	-0.388	-0.53
d_1_5	-0.662	-0.456	0.993	1.000	0.837	0.835	0.788	-0.388	-0.58
d_1_6	-0.582	-0.503	0.861	0.837	1.000	0.996	0.986	-0.371	-0.43
d_1_7	-0.582	-0.473	0.853	0.835	0.996	1.000	0.983	-0.360	-0.45
d_1_8	-0.547	-0.454	0.809	0.788	0.986	0.983	1.000	-0.370	-0.44
d_2_3	0.279	0.273	-0.388	-0.388	-0.371	-0.360	-0.370	1.000	0.677
d_2_4	0.303	-0.013	-0.532	-0.584	-0.434	-0.452	-0.444	0.677	1.000
d_2_5	0.291	0.170	-0.484	-0.488	-0.528	-0.514	-0.550	0.877	0.791
d_2_6	0.028	-0.083	-0.107	-0.150	0.320	0.312	0.374	-0.002	0.171
d_2_7	0.088	-0.043	-0.205	-0.244	0.213	0.218	0.259	0.102	0.241

By selecting pairs of parameters to view, we can rapidly isolate those parameters that separate out the cyan and blue clusters. The fragment overlay of the two clusters shows quite clearly that the C1-C8 non-bonded distance is shorter in the blue cluster than in the cyan cluster. Viewing the structure in Mercury does not suggest any significant problems with the structure.

Going through pairs of parameters like this allows us to see that the blue fragment has a shorter C1-C2 distance than the fragments in the cyan cluster, and a smaller C1-C2-C3 bond angle. However, the values of these bonded parameters are not markedly different from those in the dataset as a whole, suggesting that there is not a serious problem with the structure in the blue cluster.



In this example, the cut level at a similarity level of 0.764 has been chosen as a reasonable way to partition the data set. Additionally, the clustering at this cut-level highlights structural differences between fragments with the same basic geometry.

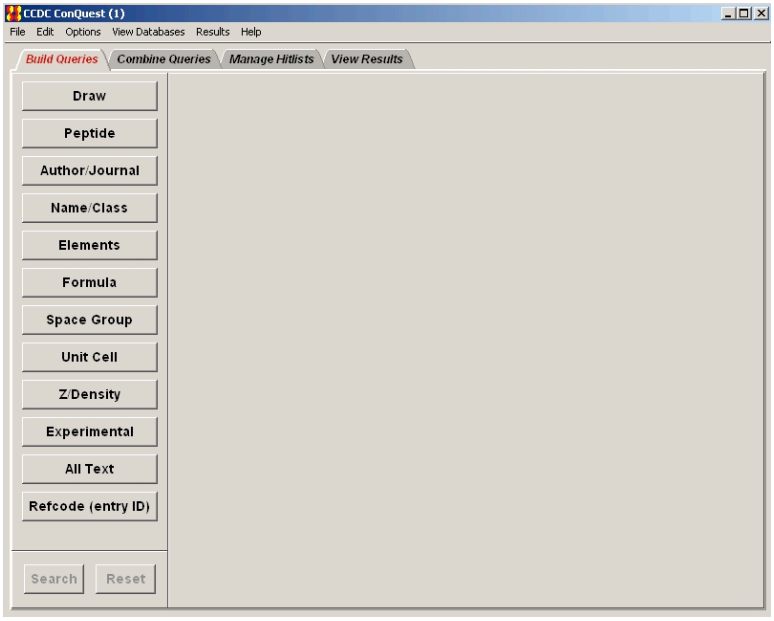
This completes Example 1.

Tip:
If you want a reminder of the search fragment, this can be viewed by pressing *F2*. This brings up a 2D chemical diagram, complete with the numbering scheme from the *ConQuest* search. It also has a list of the parameters and clicking on one of the parameters highlights it in green. This option is available in all tabs, both in Data space and in Variables space.

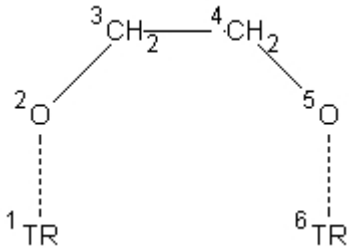
Example 2:
A simple organometallic fragment

Step 1: Set up the CSD search using ConQuest

Open ConQuest and click on the **Draw** button in the **Build Queries** tab.



Draw the fragment shown below:

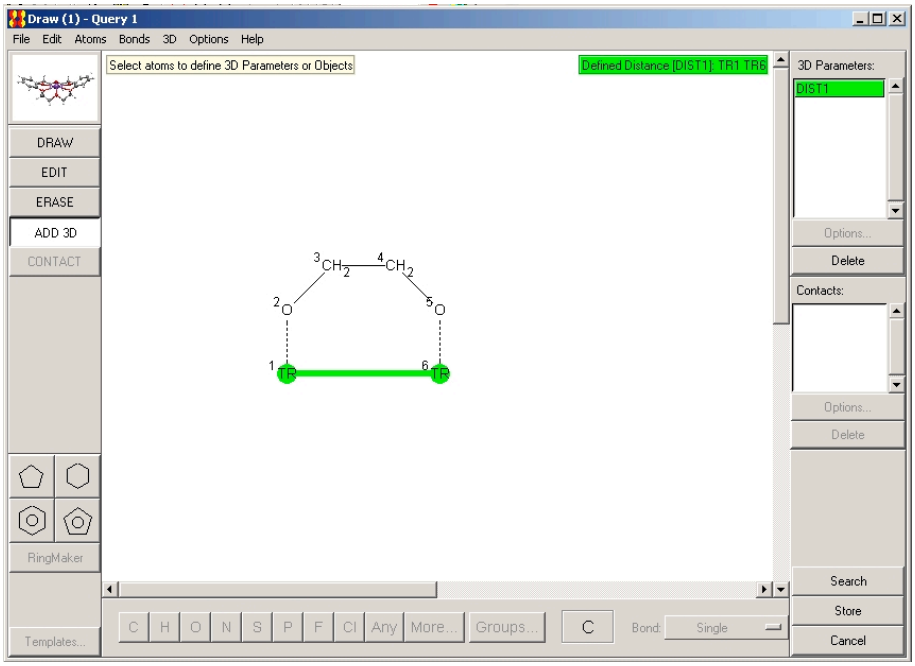


Define an interatomic distance, angle or torsion angle by clicking on the **ADD 3D** button and then selecting the atoms of interest. This needs to be done so that that the required files can be generated from the search results. It does not matter which parameter is chosen or which atoms it involves.

Tips:
The order in which the atoms are drawn corresponds to how they are numbered, and this numbering scheme is later used in *d*SNAP. For ease in following this tutorial, it is recommended that you draw the fragment to get the same numbering scheme.

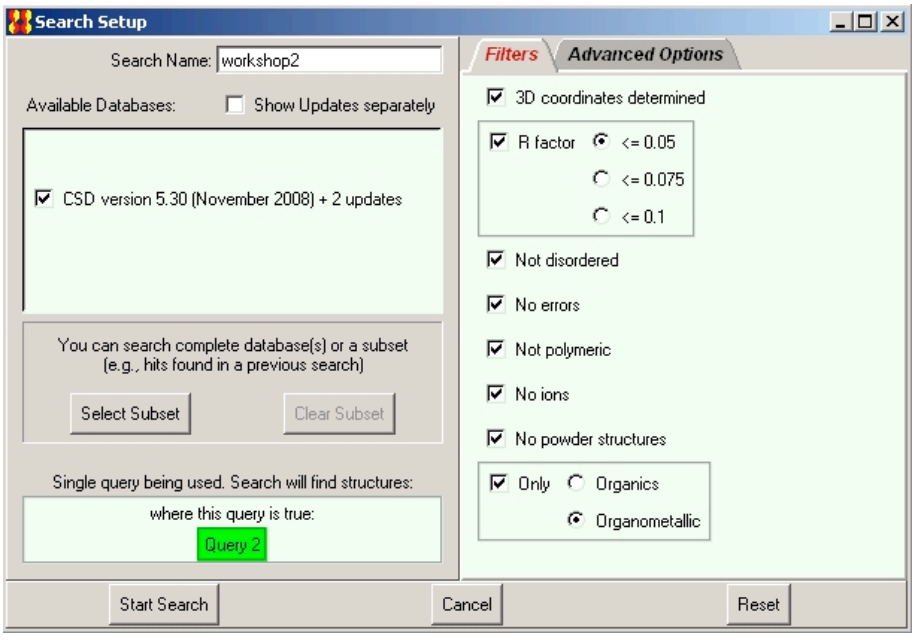
TR indicates that the element has been specified as Any Transition Metal. The dotted line between the transition metals and oxygen atoms indicates that the bond type is *Any*.

The hydrogen atoms have been defined implicitly, that is no bonds are drawn in connecting them to the carbon atoms. To do this, right-click on the atom to which hydrogens are to be added, selected **Hydrogens > Generate**. This will generate the number of hydrogens permitted by the valence of the atom. The number of hydrogens can also be specified. Alternatively from the toolbar menu choose **Atoms > Hydrogens > Generate** and click on each of the atoms that that you wish to have hydrogens on in turn. These are listed in the window. Then click ***OK***.



Tip:
If there is a parameter that you are particularly interested in, if you define that you will then have the information readily available for exporting to other programs, *e.g.* Vista or Excel.

Click the **Search** button. Change the name of the search to *workshop2*. Always make sure the ***3D coordinates determined*** box is checked in the search dialog box. For this example, also check all the other boxes, with the R-factor less than or equal to 5% and Only Organometallic.



Tip:
The maximum number of fragments that can be analysed by *d*SNAP is 4000. Note that the number of hit structures is generally smaller than the number of hit fragments, if some structures contain multiple instances of the hit fragment. It may be necessary to introduce further restrictions on the database search if the data set is large. This can be done in a number of ways, *e.g.* adding more atoms to the search fragment, including hydrogen atoms, placing stricter limits on the value of the R-factor, on experimental data collection temperature, or when the crystal structure was published.

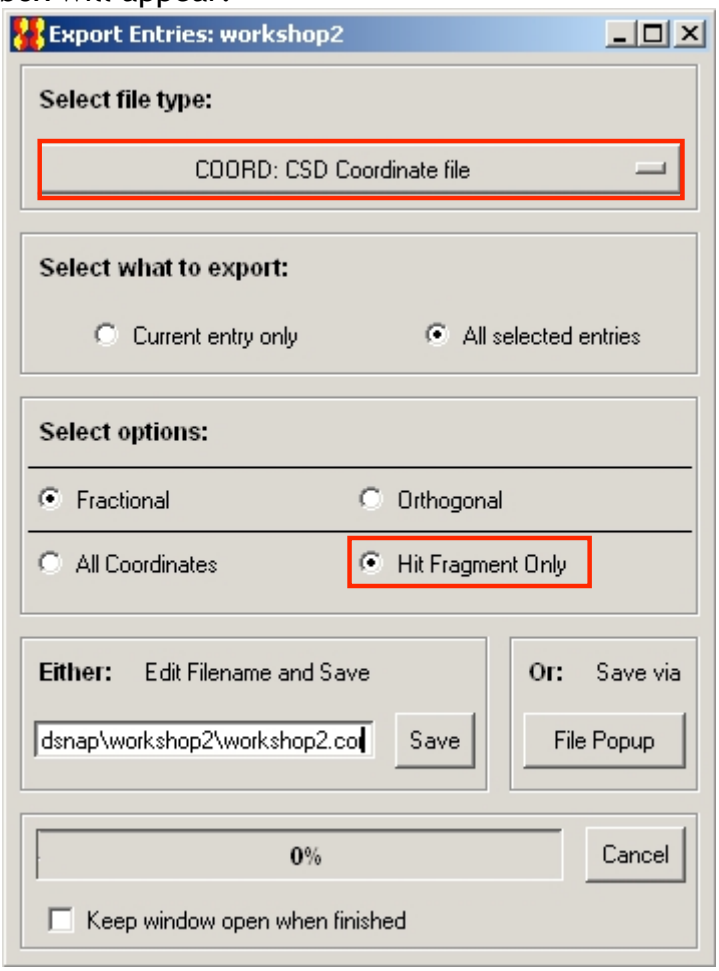
Start the search.

Step 2: Save the required files

Two files are needed for input into *d*SNAP: a .cor file, which contains a list of the coordinates of the atoms in the fragment and a .fgd file, which contains the information about atom

types, how the atoms are connected to each other and other atom and bond definition information (*e.g.* whether a bond has been defined as cyclic or acyclic, or if the connectivity of an atom has been limited). This file also contains the atom numbering scheme for the fragment.

First, create the .cor file. Go to **File > Export Entries as…** A dialog box will appear.



Select file type **COORD: CSD Coordinate file**.

Make sure that you check the box for ***Hit fragment only*** in the **Select options** section, as *d*SNAP will not run if the coordinates of all the atoms in the asymmetric unit are exported (an error message to this effect will be generated in *d*SNAP if this is done by accident).

Save this into the *workshop2* folder.

Now create the fgd file.
Go to **File > Export Parameters and Data**. Use the default options.

Note:
This folder already contains a cif, which will also be imported for inclusion in the cluster analysis.

If you are not planning on comparing your own structures to database structures, you do not need a cif.

The fgd file is created along with 2 other files, .fgn and .tab, but these are not required and can be ignored or deleted.

Make sure that all the files are saved with the same root filename (in this case, *example*2).

Keep *ConQuest* open until the *d*SNAP cluster analysis has run successfully.

Tip:
It is also strongly recommended that you save the search (as a .cqs file) for future reference.

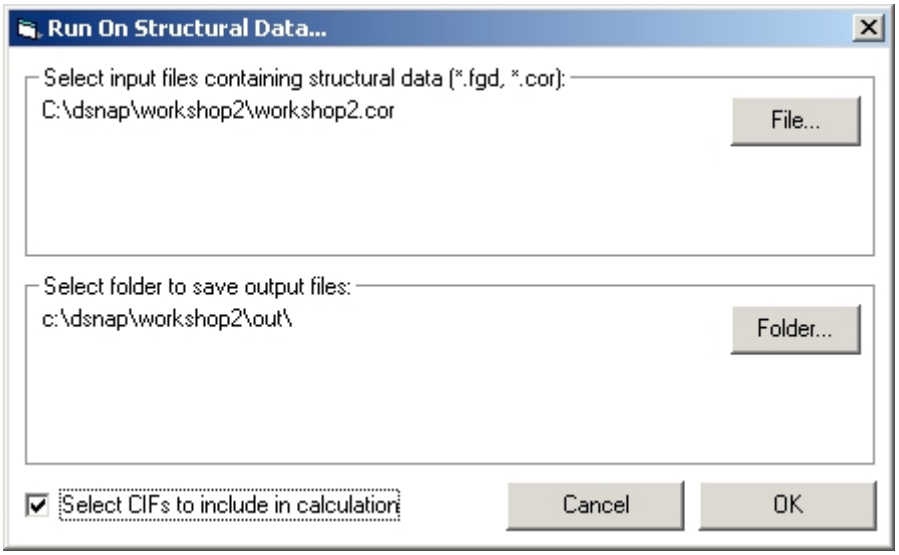
Step 3: Open *d*SNAP and identify the input files

Double-click on the *d*SNAP icon to launch *d*SNAP.

When the dialog window appears, select the **Run new analysis** button. A new dialog window will open. Browse for the correct folder containing the .cor and .fgd files and select one of them (it does not matter which).

Create the output folder. Note that the output cannot be saved in the same folder as where the input files are saved. It is recommended to create a subfolder in this directory for the output.

Check the box at the bottom left to include a cif.



Click ***OK***. A new dialogue box will appear asking for the name of the cif. Select *bca2.cif* and click ***OK***.

Notes:
If you wanted to include more than one cif, this can be done using standard multiple selection options.

Another dialog box will appear. This is used to tell the program which atoms in the cif correspond to which atoms from the original search fragment.

To associate an atom in the cif with the corresponding atom in the fgd file, right-click on the appropriate atom in the cif atom list and select the fgd atom from the drop-down list that appears. As each atom is assigned, the name of the cif atom is written (in italics) below the atom name in the diagram.

To work out which atoms are which, open the cif in *Mercury* and view the atom labels. There is only one instance of the search fragment in the asymmetric unit.

Note:
This cif is taken directly from the CSD. In this case, it helps to provide a check that all the atoms have been correctly assigned.

For this cif, the atom assignments are given in the table below:

CIF atom	Search fragment atom
Cu1	TR1
O5	O2
C7	C3
C7A	C4
O5A	O5
Cu1A	TR6

When all the atoms have been assigned, check the table on the right of the window to make sure the assignments match the list above, and then click ***OK***. You will be asked if you want to include another fragment from this cif. Click ***No***.

The cluster analysis will start. When prompted, choose ***Yes*** to perform the symmetry correction.

Tip:
This example should run quite quickly, but some analyses, particularly those involving a large number of hits from the CSD, a large number of atoms in the search fragment, or high

This fragment has topological symmetry. In this case it affects the whole fragment, but it can also affect some atoms in a fragment. In order not to produce false clusters, this symmetry must be accounted for by the program. This is done automatically by *d*SNAP for all fragments, including fragments taken from cifs.

The output for this topological symmetry correction is contained in the *pre_processing_log.txt* file, which is written to the output folder. This gives a list of the possible ways in which the atom numbering can be permuted, which are known as symmetry possibilities. For example, this fragment has two symmetry possibilities, given by:

Symm Poss	1	2	3	4	5	6
1	1	2	3	4	5	6
2	6	5	4	3	2	1

The output also lists which symmetry possibility has been assigned to each fragment. Looking at the first ten hit fragments:

```
Refcode         Symm Poss output
ALOKIV                  2
GIJKIT                  2
GIJKIT01                2
GIJKIT02                1
GIJKIT03                2
GIMKOD_01               1
GIMKOD_02               1
GIMKOD_03               1
GIMKOD_04               1
HEVSEH                  1
```

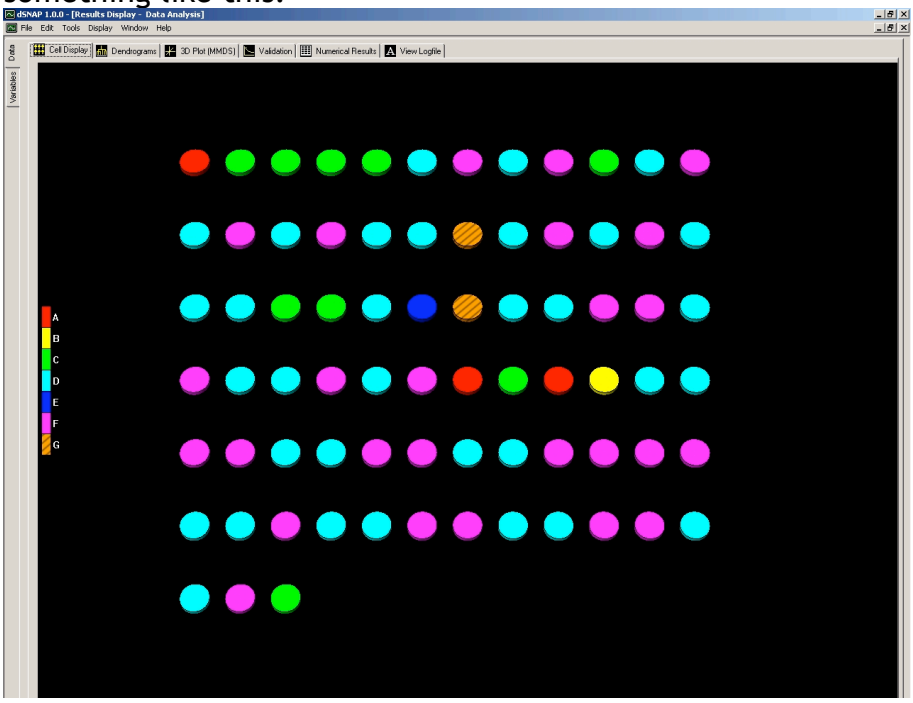
Here, those fragments with the symmetry possibility output of 2 have been renumbered, so in those fragments TR1 has become TR6, O2 is O5 and C3 is C4, while fragments with a symmetry possibility of 1 have not had their atom numbering altered.

Note:
There is a maximum limit of 4000 symmetry possibilities. In practical terms, this means that symmetry of order 7 or higher cannot be run with a symmetry correction.

Step 4: Viewing and analysing the results

Progress bars appear as *d*SNAP carries out different stages of the clustering process.

When the analysis has been completed, the first screen to appear will be the **Cell Display** in **Data** space, which will look something like this:

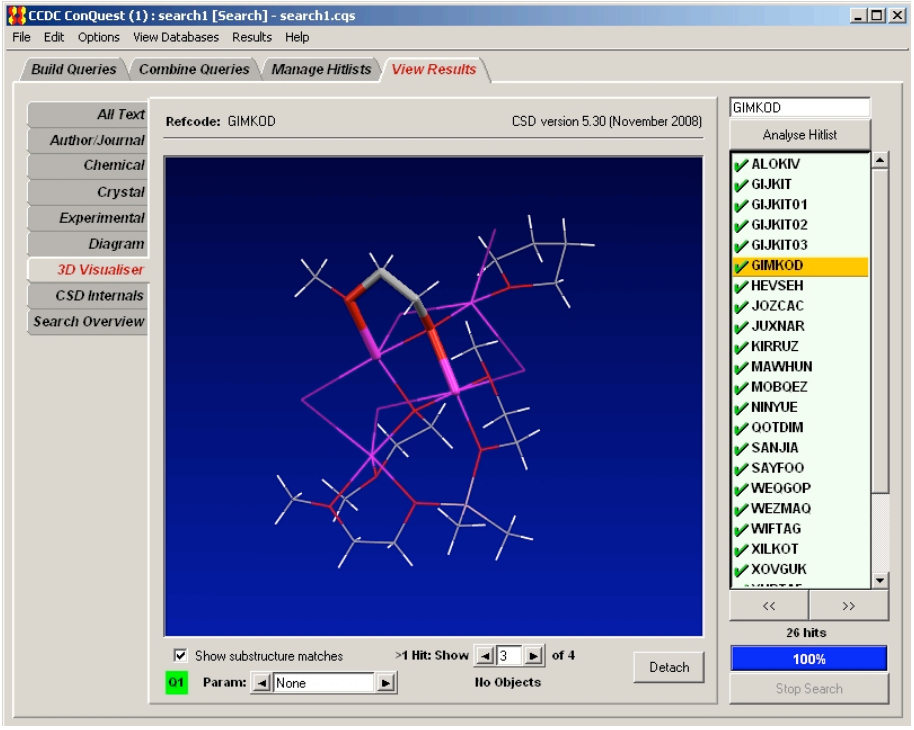


Each circle represents a hit fragment. Those which are the same colour are currently in the same cluster. The key down the left side of the screen associates each cluster colour with a letter (for up to 30 clusters; beyond this point it can be hard

to distinguish the different colours). There are currently seven clusters.

The fragments are displayed in the cell display in the order in which they are output from the CSD. This is generally alphabetically, but structures which have been included from updates to the database are found at the end. Fragments from included cif files appear last. These are also easy to spot as their identifying code will begin with a '+'.

The number of hit fragments is bigger than the number of hit refcodes from the CSD search. This is because the search fragment occurs more than once in some of the structures, due to multiple instances of the same fragment in a single molecule, or because of multiple molecules, *Z'*>1, in the asymmetric unit of the structure. When this occurs, the refcode has *_nn* appended to it, where *nn* is the number of the fragment in that structure. If you refer back to the original CSD search you can highlight these hits individually, *e.g.* refcode GIMKOD, which has 4 sub-structure matches for the search fragment, labelled GIMKOD_01, GIMKOD_02, GIMKOD_03 and GIMKOD_04.



Tip:
Here, the third hit fragment is being highlighted as capped sticks, while the rest of the structure is shown as wireframe in the *ConQuest* **3D Visualiser**. In *d*SNAP, this hit will correspond to GIMKOD_03. This display option is accessed by right-clicking in the *ConQuest* display window and picking **Highlight Hit > Hit as Capped Stick.**

Fragments in the Cell Display in *d*SNAP can be selected by clicking on them. Multiple fragments can also be selected. If the fragments are in sequence, click on the first, hold down the *Shift* key and click on the last. If the fragments to be selected are not in sequence, hold down the *Ctrl* key and click on each of the fragments of interest. An entire cluster can be selected by clicking on the coloured box corresponding to the

cluster in the key on the left of the screen. Multiple clusters can be selected by clicking on the each cluster and holding down either the *Shift* or the *Ctrl* key. Note each cluster must be selected individually.

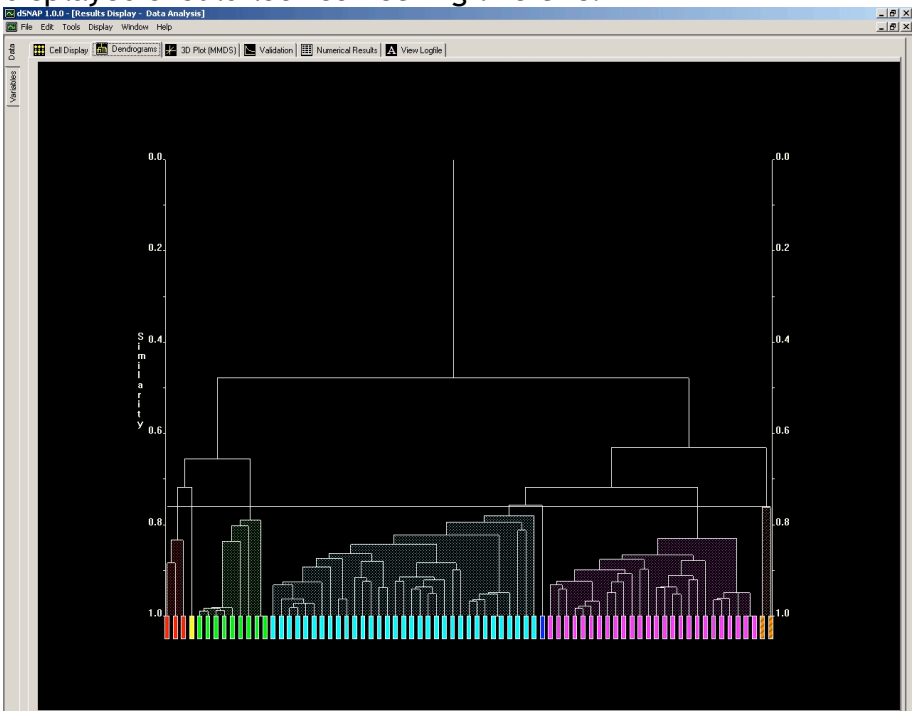
The selected hits can be viewed in Mercury, either by pressing F3, or through the menu. *Go to* ***Tools > Show Selected Hits in Database Viewer… F3***.

Each fragment has a label. By default these are not shown for datasets of more than 15 fragments. However, by right-clicking anywhere in the display, the labels can be shown for all fragments (***Show All Labels***) or for selected fragments (***Show Selected Labels***).

Note:
These options for viewing labels are also available in the other graphics tabs.

In *d*SNAP, how the clusters are assigned can be seen and adjusted by viewing the dendrogram.

Click on the **Dendrogram** tab. The dendrogram being displayed should look something like this:



Each box at the bottom of the screen represents a single hit fragment. The boxes are joined by horizontal lines called tie bars, which link together fragments according to the calculated similarity between each connected branch. The vertical axis is a similarity scale, with zero similarity at the top, and a similarity of 1.0 at the bottom *i.e.* if two fragments are joined by a tie-bar near the bottom of the dendrogram then they can be considered very similar, justifying their being grouped together. If two branches do not meet until near the top of the dendrogram, the associated fragments

Notes:
For datasets with a very large number of hit fragments the individual boxes may not be distinguishable and you may need to zoom in to see the finer details.

Because the imported cif was taken from a CSD structure which is also included in this analysis, it can be identified as it is linked to another fragment with a similarity of 1.

are only loosely related to each other.

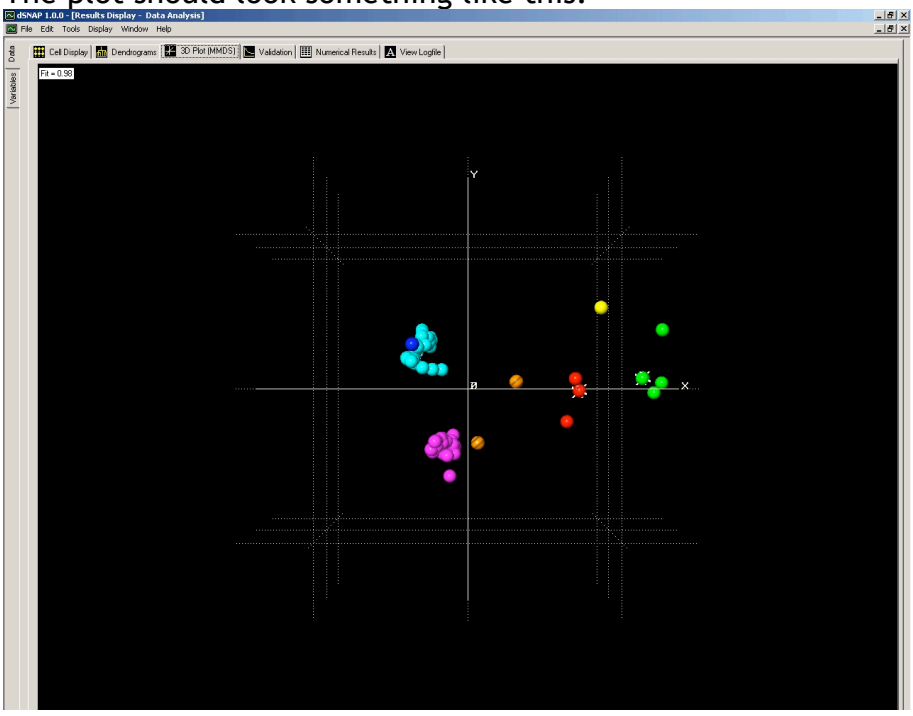
The horizontal line that spans the width of the dendrogram marks the cut level. Any fragments that are linked at higher level of similarity than the cut level (*i.e.* lower down the dendrogram) are put into the same cluster), while those which are less similar (*i.e.* their tie bars lie above the cut level on the dendrogram) belong to separate clusters. The current cut level is 0.76. The initial cut level is determined using Principal Components Analysis, but can be adjusted by left-clicking anywhere in the dendrogram display and scrolling using the scroll wheel on the mouse.

Note:
Alternatively, if you do not have a scroll wheel on the mouse (*e.g.* using a laptop) the cut level can be adjusted by holding down the *Alt* key, left click on the cut level and drag the cut level up or down). It can also be adjusted via the menu: **Tools > Set Cut-Level To…** and then enter the required value. This also allows the cut-level to be set to a specific value.

Moving the line up (raising the cut level) will decrease the number of clusters, and overall the fragments within each cluster will be less similar to each other, while lowering it will increase the number of clusters, and the fragments in each cluster will be more similar to one other.

See the effect of raising and lowering the cut level on the numbers of clusters. Then click on the **3D Plot (MMDS)** tab. At this stage, do not save the changes to the dendrogram.

The plot should look something like this:



Each sphere represents a hit fragment. Fragments that are located close together in space can be expected to have similar geometries. The colours are taken from the dendrogram. Ideally, spheres of the same colour will be close together and well separated in space from spheres of different colours. This indicates good agreement between the MMDS

plot and the dendrogram. The plot can be rotated in space by dragging while holding down the left mouse button.

Some of the spheres have white spikes coming from them. These indicate the most representative fragment in each cluster; this can be useful for selecting examples to illustrate the geometries of each cluster. They are only shown for clusters containing three or more fragments.

Note:
The most representative sample highlighting can be switched off by right-clicking anywhere within the 3D Plot display and clicking on ***Show MRM Marks***.

Under the Validation tab are several tools for analysing results. These are not so relevant to this tutorial, but are explained in the *d*SNAP manual.

In order to assess whether the default cut level is the most appropriate, first compare the geometries of the fragments within each cluster to see whether they are similar enough* to be included in the same cluster. Then compare the clusters to one another.

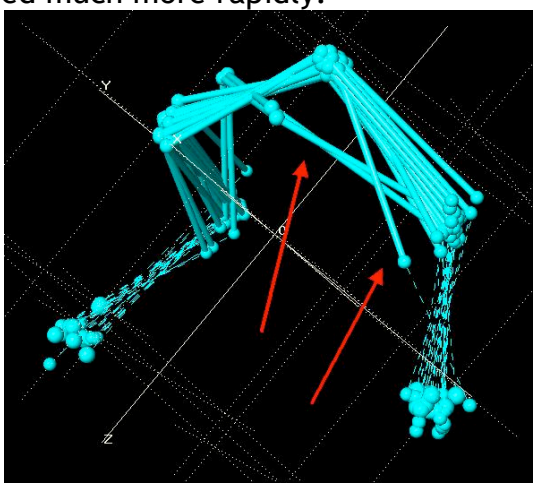
*** 'Similar enough' is subjective, and very much dependent on the level of detail that is relevant and appropriate to the investigation. At the two extremes, all fragments could be considered in a single cluster as they all have the same connectivity, or all fragments whose similarity is less than 1 are different as they do have different geometries, no matter how small the differences are. For most datasets, a situation somewhere between these is probably the most appropriate or meaningful, and it is the justification of cut level that is important. The initial cut level provides a good starting point for the number of clusters in the data set, but it may need to be adjusted depending on the level of detail that is required from the analysis.**

One of the things to note is that in the MMDS plot several of the groups are quite diffuse. In particular the two largest clusters (coloured cyan and pink) appear to be split.

Switch to the Dendrogram tab. Highlight all the fragments in the cyan cluster and press *F1*. This brings up the **Multiple Fragments Viewer**, which displays the best overlay of all the selected fragments.

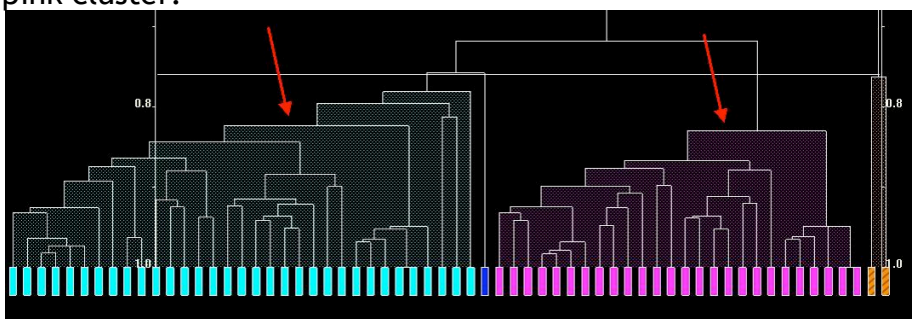
Tips:
An easy way to do this is to switch to the Cell Display, left-click on ***D*** in the key. This selects the whole cluster. Then switch to the dendrogram display. They do not appear selected, but pressing *F1* will bring up the fragment viewer for the whole cluster.
The Multiple Fragment Viewer can also be accessed through the menu. Go to **Tools > Show Selected Fragments in 3D Viewer... F1**.

The fragment viewer is a very useful tool in establishing the suitability of a chosen cut level as it provides an excellent check of how similar the fragments are when viewed en masse. Small differences that may be lost when comparing the structures individually can become much more apparent. This allows outliers for a given cluster, or the dataset as a whole, to be identified much more rapidly.

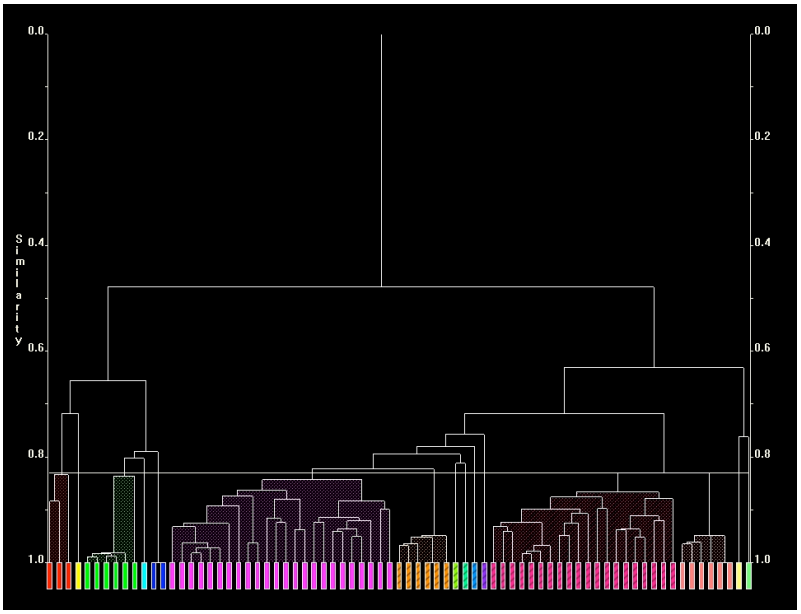


Although the majority of the fragments appear to be overlaying well, there are a few which have a slightly different geometry (these are marked with arrows in the above figure).

Close the fragment viewer. Looking at the dendrogram, there is quite a big step between two groups of fragments in the cyan cluster. This is a good indication that it may be appropriate to lower the cut level and split the cluster further into two groups. A similar situation is also observed in the pink cluster.

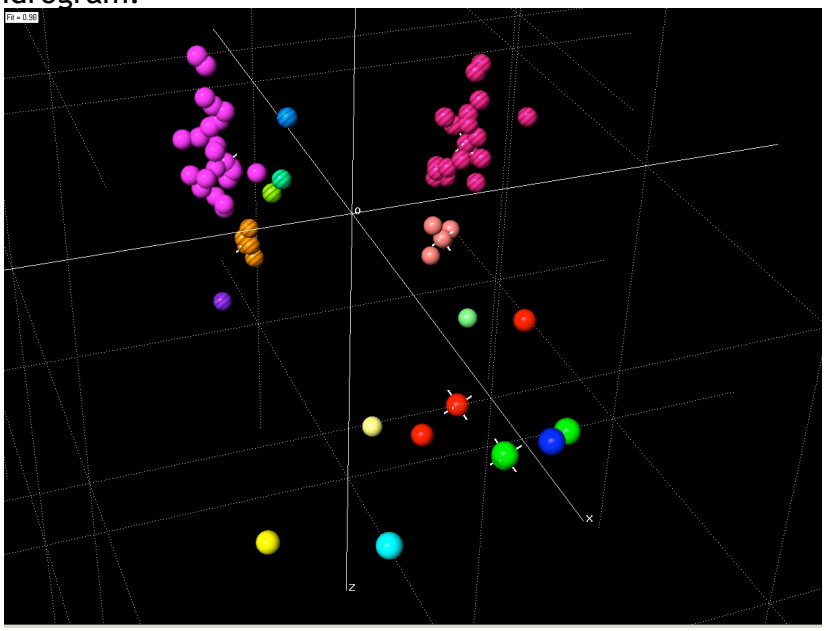


Lower the cut level to around 0.83. The dendrogram will look like this:



Notice that changing the cut level affects all the clustering colours. In this instance, several new clusters have been created and there are now 15 clusters.

Click on the 3D Plot tab, saving the changes to the dendrogram.



Note:
This image shows a zoomed-in and rotated view of the MMDS plot, with the sphere size decreased.

Sphere size can be adjusted by holding down the left mouse button and dragging up or down on the screen while pressing *Alt*.

Notice that the groups of fragments that appeared to be grouped separately at the old cut level are now in different clusters.

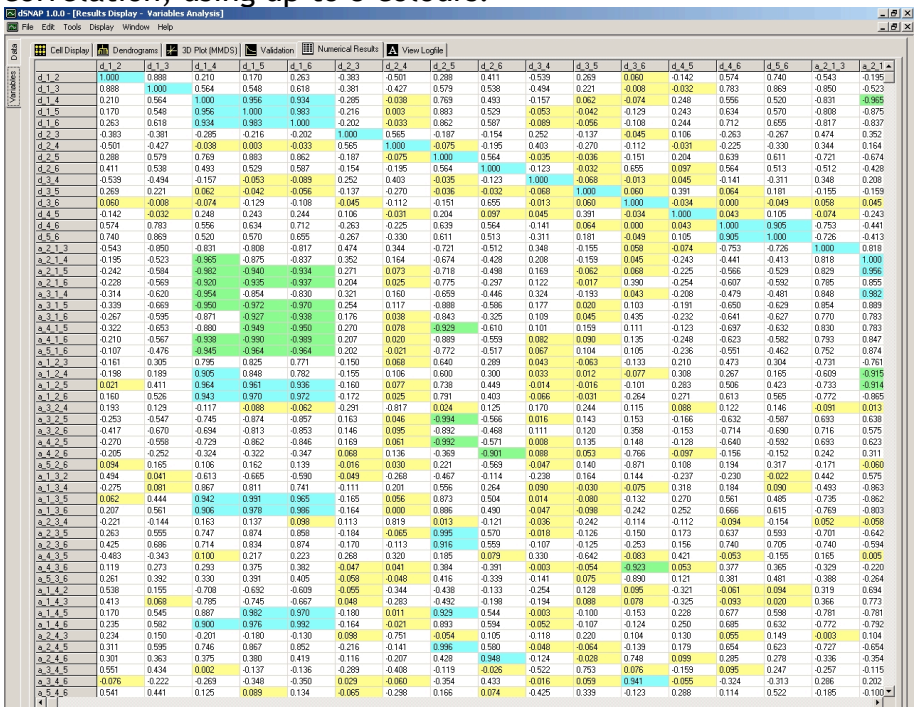
Go back to the Dendrogram tab and view each cluster in the Multiple Fragments Viewer to verify that the groupings make sense.

Viewing the pink and brown striped clusters at the same time with the Multiple Fragments Viewer, it appears that the two

groups have almost identical geometries, but the MMDS plot suggests that they are separate clusters. It is therefore necessary to identify which parameters are important in separating these fragments into two clusters.

Go to the **Numerical Results** tab and then click on the **Variables** tab on the left of the screen.

This brings up a matrix giving the correlations between all possible pairs of parameters (the distances and angles), ranging between –1 to +1. The diagonal is always +1. The matrix can be colour-coded for different values of the correlation, using up to 5 colours.



We will use the Numerical Results to compare the bonded distances and the bond angles. Clicking on a cell brings up a scatterplot relating the two parameters.

In the plot, each dot represents a fragment. They are colour-coded according to cluster, as determined by the choice of cut level in the last saved dendrogram. Hovering over the fragment brings up a tool-tip box with the name of the fragment and the value of each of the two parameters for that fragment. Clicking on a dot selects it and opens *Mercury* to display it.

Notes:
Viewing the Numerical Results in Data space displays the correlation matrix between each fragment.

The choice of colours and the range for each colour can be changed in the **Edit > Options** menu. Click on the ***Select Correlation Matrix Colours*** box at the bottom left of the window. This brings up a new window headed ***Numerical Results Formatting***. Left-clicking on the colour boxes brings up a colour selection box. The use of dark colours is not recommended as the values of the correlations will become difficult to read. The colours will be up-dated the next time that the Numerical Results are viewed after switching from Data space.

Including the colours means this tab takes longer to load, and so it is switched off by default if there are more than 500 parameters (this corresponds to a 10 atom fragment). You can increase this value if you wish to colour-code the correlation matrix for larger fragments. The table gives the number of parameters for different numbers of atoms in the fragment, up to the program's maximum limit of 20 atoms.

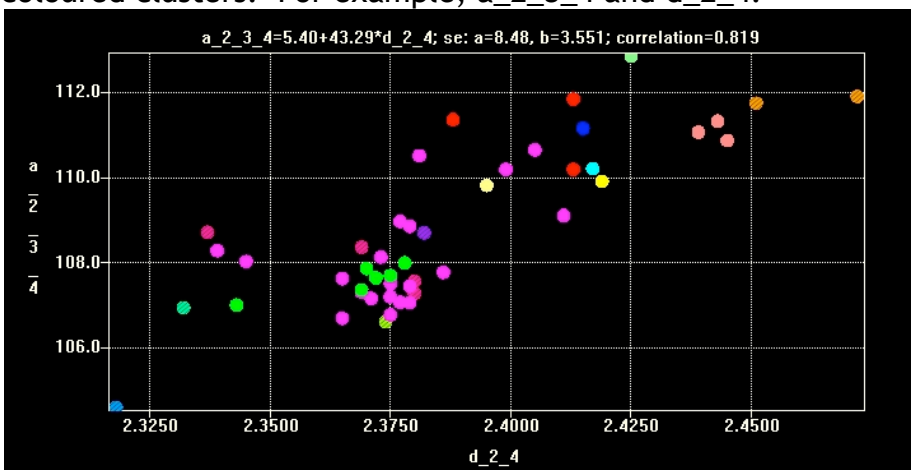
Number of atoms	Number of parameters
11	550
12	726
13	936
14	1183
15	1470
16	1800
17	2176
18	2601
19	3078
20	3610

So if we wish to view the relationship between the TM1-TM6 distance and the TM1–O2 distance, we could click in either of the cells highlighted below:

	d_1_2	d_1_3	d_1_4	d_1_5	d_1_6	d_2_3	d_2_
d_1_2	1.000	0.888	0.210	0.170	0.263	-0.383	-0.50
d_1_3	0.888	1.000	0.564	0.548	0.618	-0.381	-0.42
d_1_4	0.210	0.564	1.000	0.956	0.934	-0.285	-0.03
d_1_5	0.170	0.548	0.956	1.000	0.983	-0.216	0.00
d_1_6	0.263	0.618	0.934	0.983	1.000	-0.202	-0.03
d_2_3	-0.383	-0.381	-0.285	-0.216	-0.202	1.000	0.56
d_2_4	-0.501	-0.427	-0.038	0.003	-0.033	0.565	1.00
d_2_5	0.288	0.579	0.769	0.883	0.862	-0.187	-0.07
d_2_6	0.411	0.538	0.493	0.529	0.587	-0.154	-0.19
d_3_4	-0.539	-0.494	-0.157	-0.053	-0.089	0.252	0.40

Tip:
If you want a reminder of the search fragment, this can be viewed by pressing *F2*. This brings up a 2D chemical diagram, complete with the numbering scheme from the *ConQuest* search. It also has a list of the parameters and clicking on one of the parameters highlights it in green. This option is available in all tabs, both in Data space and in Variables space.

By selecting pairs of parameters to view, we can rapidly isolate those parameters that separate out the pink and brown striped clusters, and also the pink-striped and terracotta coloured clusters. For example, a_2_3_4 and d_2_4:



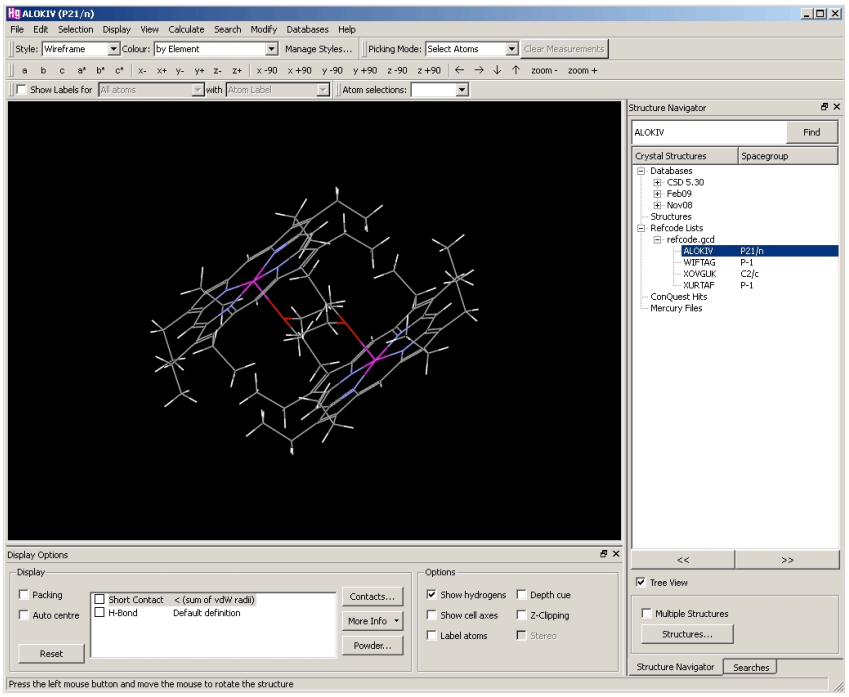
Tip:
The size of the dots has been increased for clarity. To do this, hold down the left mouse button and drag up or down on the screen while pressing Ctrl.

We find that, in both cases, there is a difference in the values of the O2-C4 distance, with a corresponding difference in the O2-C3-C4 angle.

Having established that the geometries of all the different clusters are in some way distinct from one another, it is now necessary to verify that the clusters make chemical sense. This can be done by viewing not just the fragment but the whole molecule, or if necessary, the entire unit cell. This can easily be done using *Mercury*.

Highlight the fragments in the red and yellow clusters and view the geometries in the Multiple Fragments Viewer. Keep the group highlighted and press *F3*. This opens *Mercury*.

Note:
The option to view the complete structure is also available through the menu. Go to **Tools > Show Selected Hits in Database Viewer… F3**. This option can be also be accessed from the Cell Display and 3D plot in Data space, and when a fragment is selected in a scatterplot generated in the Numerical Results tab in Variables space.



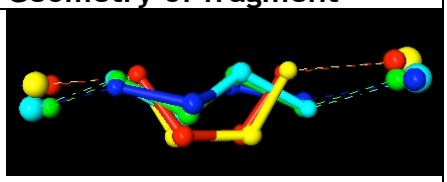
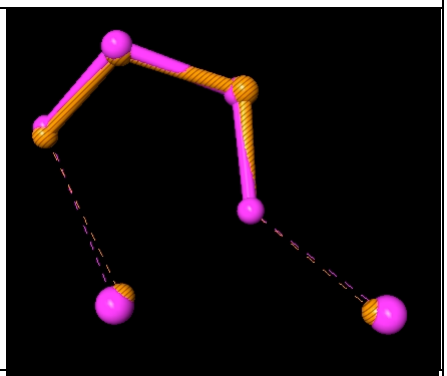
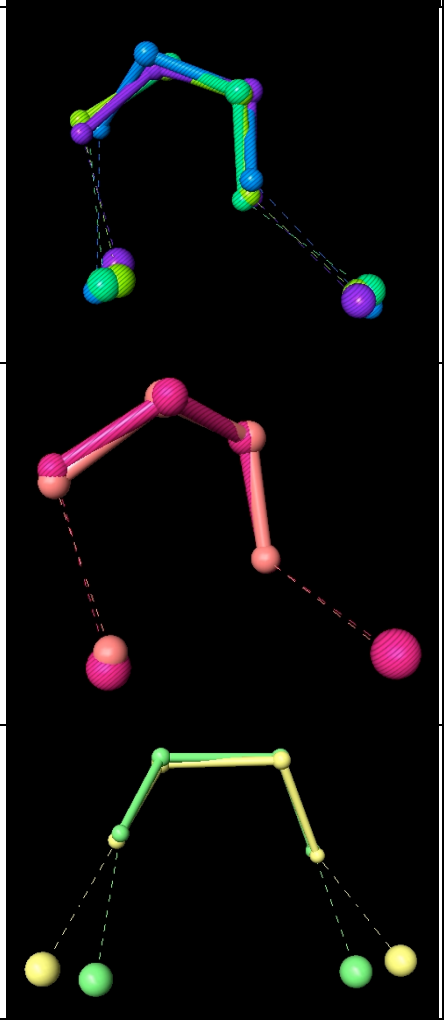
All the refcodes for the structures in the cluster are listed. Look at each in turn. In three of the four structures, the oxygen and carbon atoms are found in a 1,4-dioxane ligand. All structures have the two transition metals in a *trans* arrangement.

Metal type is also found to be an important in affecting the geometry. In the pink cluster, the metal is Mn, Ti or Cd, while in the brown striped cluster it is Co, Ni or Fe. The same pattern is observed between the pink-striped and terracotta clusters. In fact, both sets of clusters correspond to the corresponding hit fragments.

It is then interesting to ascertain why some hit fragments in a single hit structure are in one cluster and others are in a different cluster. In the case of the GIMKOD, for example, it relates to a difference in the relative positions of C3 to TM6.

In this example, the cut level of 0.83 has been chosen as a reasonable way to partition the data set as it separates the fragments on their different geometries, and also shows discrimination in the chemistry of the fragment, but distinguishing some different types of transition metal atoms.

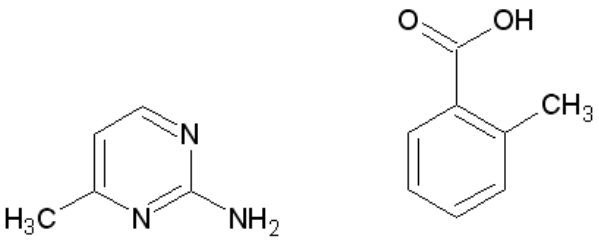
The classification of the clusters is summarised in the following table.

Cluster (colour)	No. of fragments	Geometry of fragment
A (red)	3	
B (yellow)	1	
C (green)	6	
D (cyan)	1	
E (blue	2	
F (pink)	24	
G (brown stripe)	6	
H (green stripe)	1	
I (mint green stripe)	1	
J (blue stripe)	1	
K (purple stripe)	1	
L (pink stripe)	20	
M (terracotta)	6	
N (light yellow)	1	
O (light green)	1	

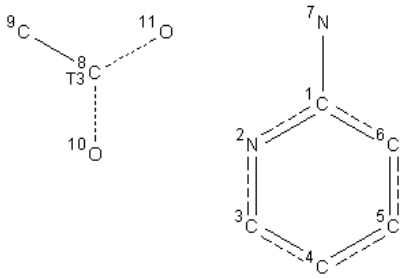
This concludes Example 2.

Example 3:
Investigating intermolecular interactions

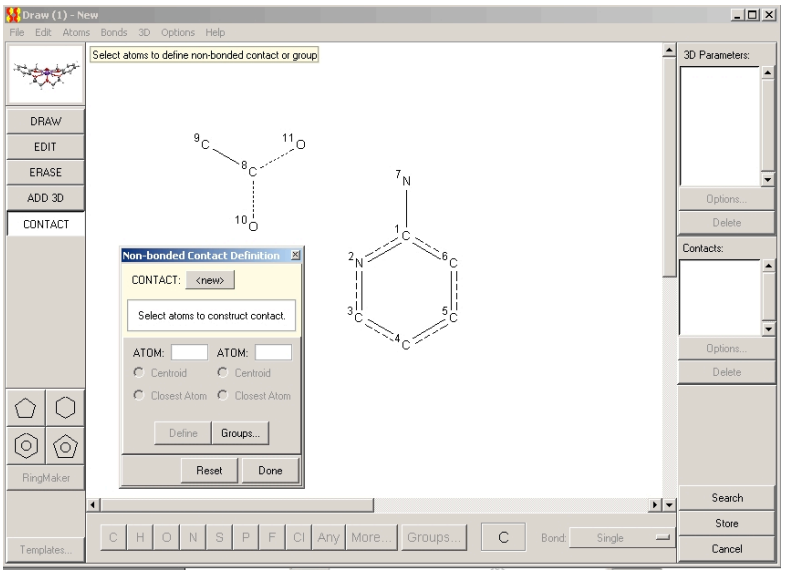
The search structure is inspired by the co-crystal in the 2007 CCDC blind test:



Draw the two fragments in *ConQuest* as shown below:



Click on the **CONTACT** button. This brings up the **Non-Bonded Contact Definition** window.



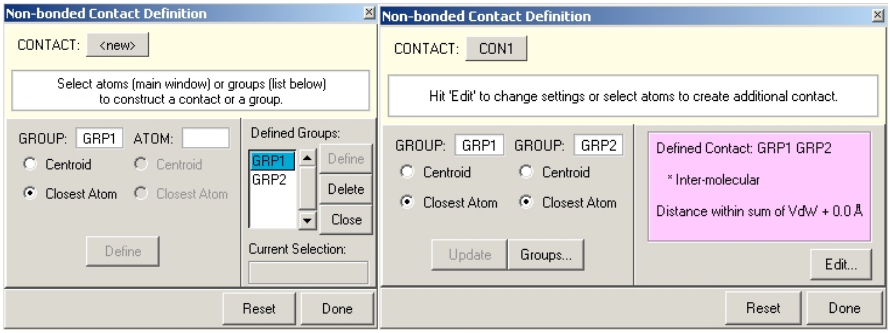
In this case we want to define a contact between two groups of atoms, the three atoms N2, C1 and N7, and the three atoms C8, O10 and O11. First these groups of atoms need to be defined. Click on the **Groups…** button in the window. Click on the three atoms forming one group and click **Define**.

Notes:
The bond type for the C8-O10 and C8-O11 is *Any*, and C8 has been defined to have three bonded atoms (T3).

The way the fragment is defined means that the carboxylate component of the fragment possesses local symmetry (atoms O10 and O11 are equivalent), so a symmetry correction should be applied prior to cluster analysis.

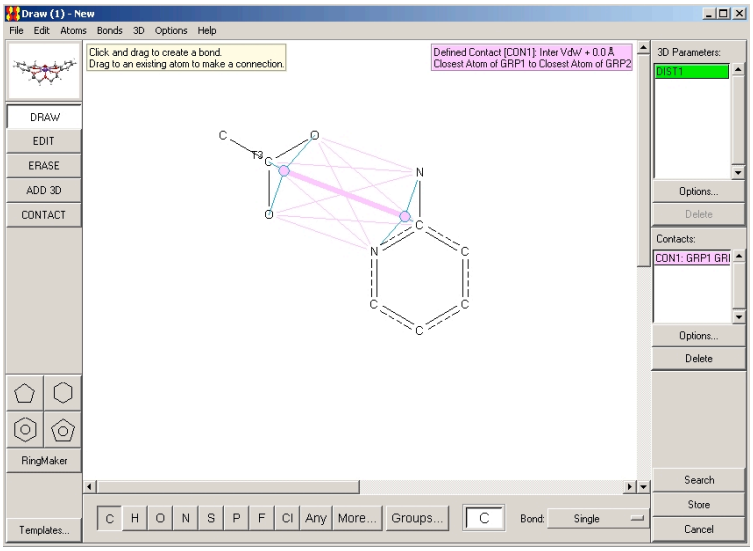
Note:
As you click on each atom it is listed in the **Current Selection** box and is highlighted (in pink when there are 1-2 atoms selected, then in blue, with lines linking the group, for 3+ atoms). When a group has been defined it appears in the **Defined Groups** list, in the form GRP1, GRP2, *etc*.

When both groups have been defined, click on each group name in the **Defined Groups** list in turn. They will appear on the left side of the window. Check the **Closest Atom** box for both groups. Then click the **Define** button.



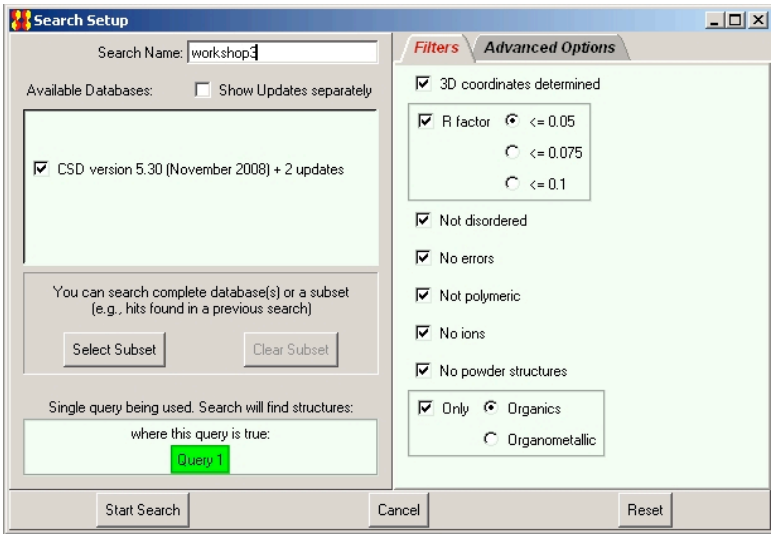
By default the distance between the closest atoms is the sum of the van der Waals radii.

Click **Done**.



Note:
It is not necessary to define a parameter to be able to produce the required output for *d*SNAP as setting up the intermolecular contact defines a distance; a **3D Parameter** is listed, DIST1.

Start the search, changing the search name to *workshop3* and selecting all the search filters as show
n:

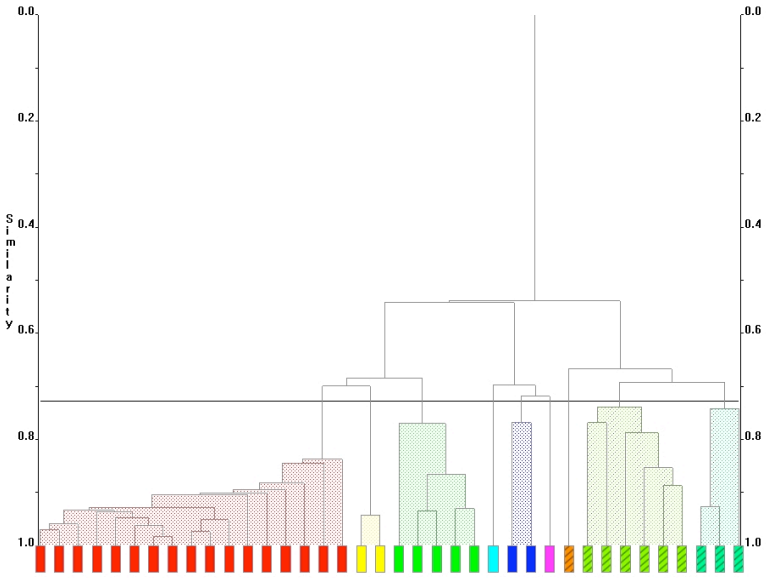


Note:
Because of the number of atoms in the fragment, the cluster analysis may take longer than the other workshop datasets even though the data set is small.

Save the cqs, cor and fgd files into the *workshop3* folder with the root filename *workshop3*.

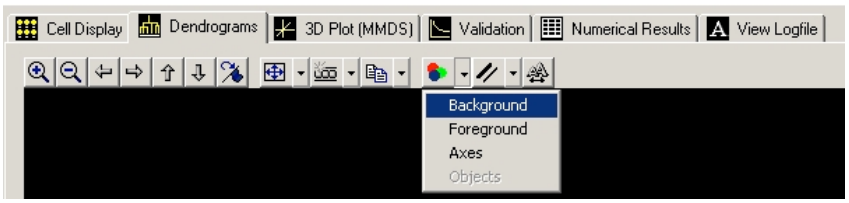
Start the *d*SNAP analysis. When prompted, choose to perform the symmetry correction.

The dendrogram will look something like this:



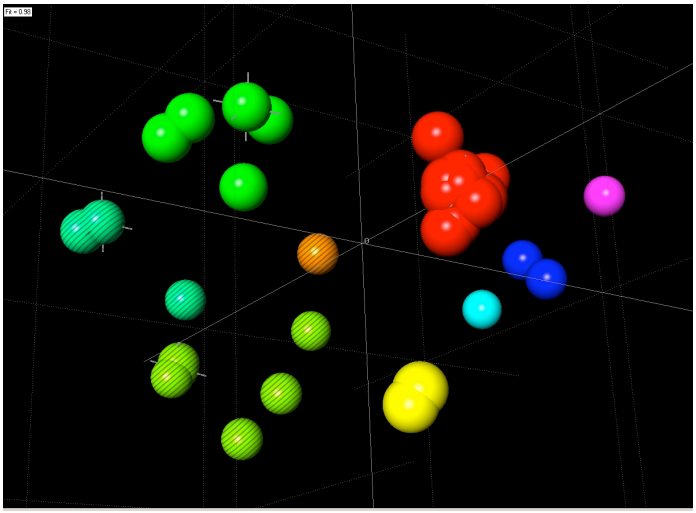
Notes:
Here the background and axes colours of the dendrogram have been changed to produce a picture that has been optimised for use in publications. The white background reduces the ink requirements. ***Axes*** (which includes the cut level) have been coloured black, and the ***foreground*** colours (which include the tie bars and any fragment labels) have been coloured dark grey for contrast with the cut level.

To change the colours, right-click in the display area, and select **Show Toolbar** from the drop-down menu. When the Toolbar is shown it appears at the top of the display like this:



Note:
In the toolbar, click on the arrow by the **Change color** button. This brings up the list of items whose colours can be changed. Selecting an item brings up a colour picker box. Select the colour required and click ***OK***.

Look at the MMDS plot:



Note:
MMDS plots are sharper when produced on a black background, both on screen and in print. The picture quality can be improved by increasing the rendering quality. Click anywhere in the plot area and press *F12* to bring up a dialog box. The rendering quality is normally set low, to speed up rotation when dealing with large datasets, but increasing the quality will give smoother lines and better results when printed.

Now view each cluster in turn at the default dendrogram cut level using the Multiple Fragments Viewer.

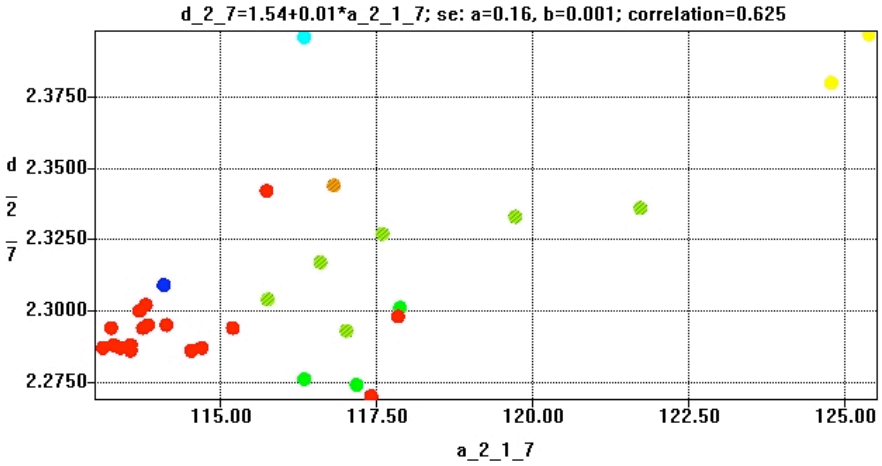
Try altering the cut level. Ideally, you would like to find a cut level where the geometries of all the fragments are similar within the cluster, but distinct from the geometries in other clusters. Remember that you make your own judgement on how similar fragments within a given cluster should be to one another, as long as you are prepared to justify your choice. You may decide that the default cut level is best, but in any analysis you should always investigate the effects of raising and lowering it.

We will take the default cut level of 0.728 as the starting point.

Viewing the red and yellow clusters together in the multiple fragments viewer, it appears that they have very similar geometries; both involve a dimeric relationship between the two groups, with interactions between N7 and O10, and N2 and O11. However, the two groups are distinct in the MMDS plot.

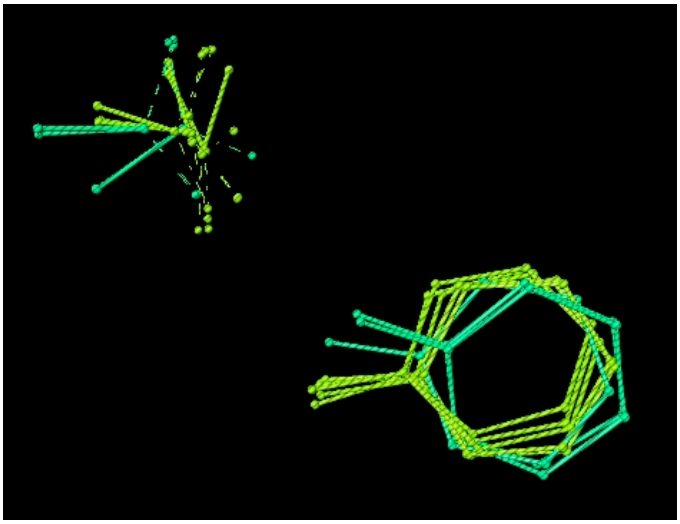
Use the scatterplot facility in the Numerical Results in Variables space to investigate the differences in the bonding distances and angles between these two clusters.

We find that these two clusters differ in the geometries of the 6-membered ring. Now view the structures in the yellow cluster in *Mercury*. In the yellow cluster, N7 is part of a 5-membered ring fused to the 6-membered ring of the search fragment. Look at the N2-C1-N7 angle in the Numerical Results (compare it to any other parameter); this angle is larger for the yellow cluster than for any of the other clusters.



Because of this, we might want to choose a cut level which maintains these two sets of fragments as separate clusters.

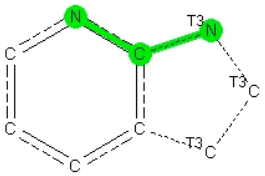
Compare the two green striped groups with the multiple fragmen



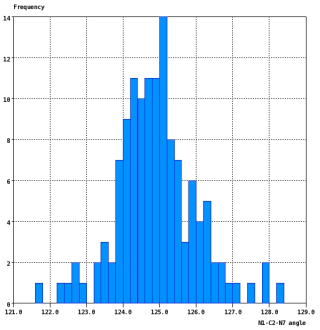
Close the viewer window, then, without deselecting the fragments, press *F1* to bring up the viewer again. This time when the window with the chemical diagram appears, deselect all the atoms in the carboxylate group by clicking on them (they will no longer be

Note:
By default, no colours are shown in the Numerical Results matrix display because of the number of atoms in the search fragment.

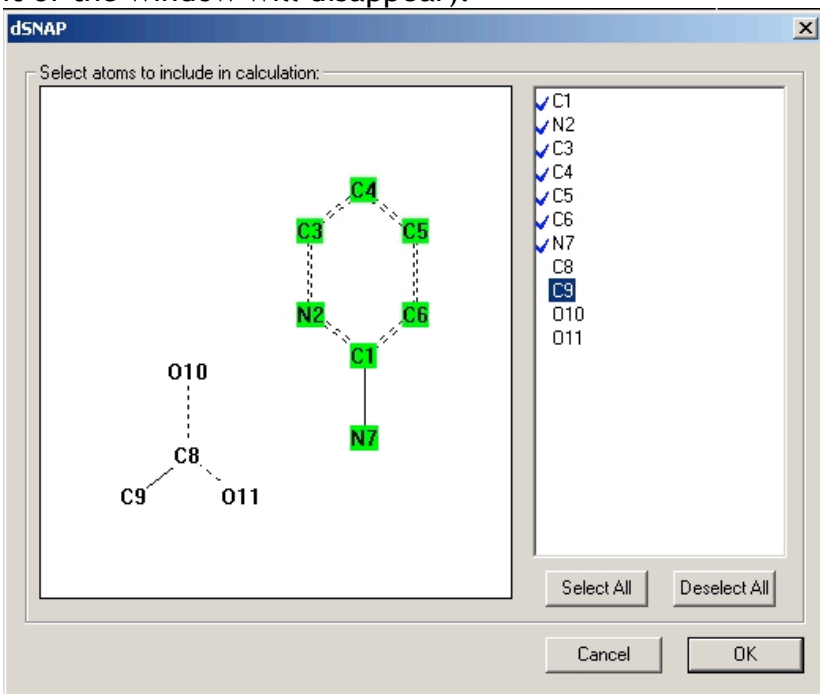
Note:
A search for this fragment in the CSD, defining the N2-C1-N7 angle, shows that the size of the angle is within the normal range for this type of structure:



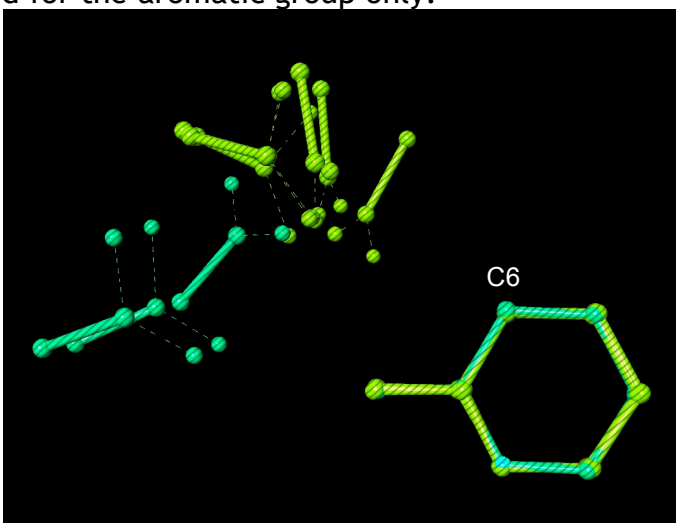
This information is rapidly obtained using Vista:



highlighted in green, and the tick mark next to them in the list on the right of the window will disappear).

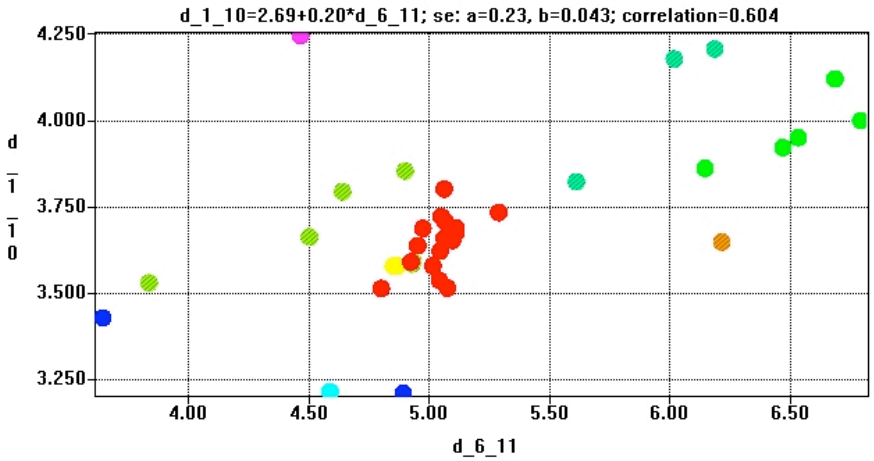


This produces a figure where the overlay of the fragments is optimised for the aromatic group only.



Note:
The display of gridlines and axes in the Multiple Fragments Viewer and the 3D Plot can be switched off by right-clicking in the display and deselecting the

This makes the differences between the two groups clearer. The olive-striped group is positioned further towards C6 than the mint-striped group. This is reflected in the scatterplot relating C1 and O10, and C6 and O11:



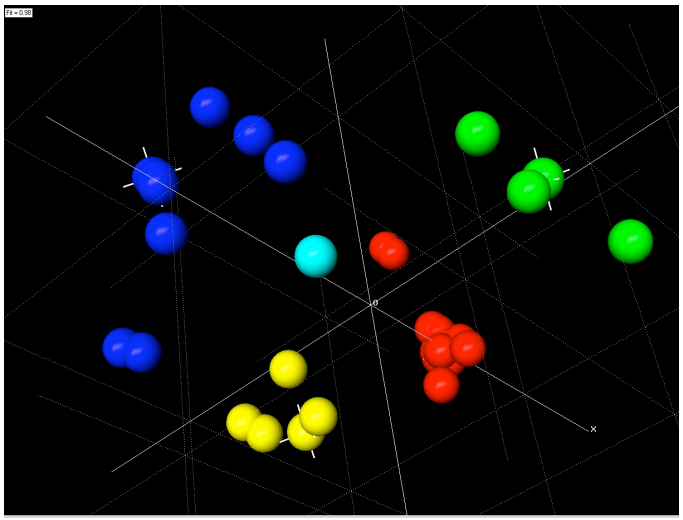
The cyan, blue, pink and brown-striped clusters each contain only 1 or 2 fragments. Viewing them together shows that they all have different geometries, but are similar to one another as they all involve inter-planar contacts.

Examining the green cluster shows that it also involves an interaction between N7 and O10, but there is no interaction involving N2 and O11, and so it differs from the red and yellow clusters.

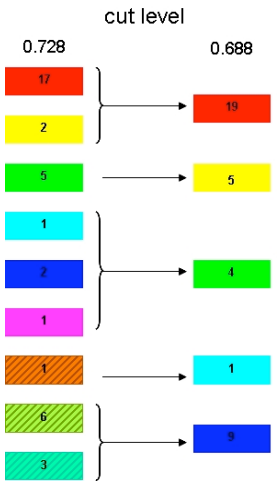
In this data set, the red cluster is very compact in the MMDS plot and has the lowest tie-bars in the dendrogram, which reflects that the nature of the dimeric interaction limits the variation in the intermolecular distances and angles to a greater extent than the other interactions, but other clusters are quite diffuse.

Note:
By the nature of inter-molecular interactions, there tends to be greater variation in the overall geometries of the groups and clusters are often more diffuse than is the case with intramolecular geometries.

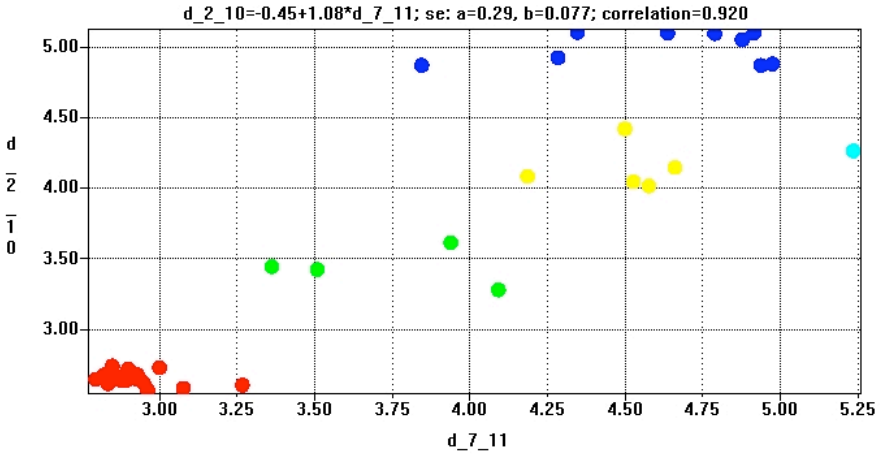
Another interpretation therefore might be to raise the cut level to 0.688. There are now five clusters. The MMDS plot shows that the clusters are still distinct from each, although they are less well-defined than at the higher similarity value of the default cut level.



Note:
The following diagram shows the relationship of the clusters at the two cut levels:



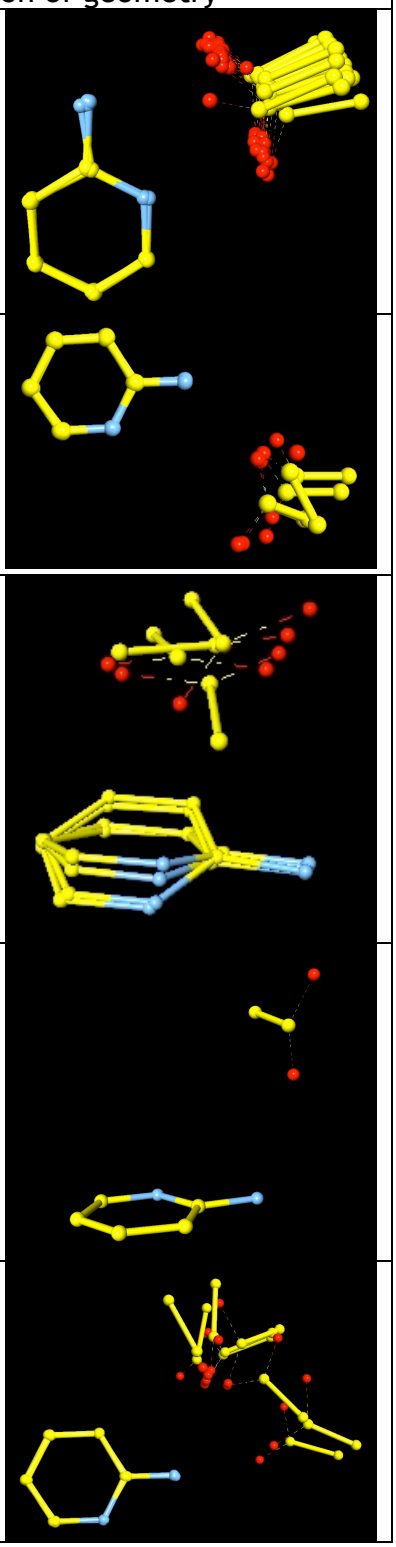
By examining the interatomic distances and angles, at this choice of cut level we can see that there is good discrimination between the clusters on the basis of just two parameters, the N2···O10 distance and the N7···O11 distance:



Examining the clusters in the Multiple Fragments Viewer shows that these clusters reflect the main dispositions of the carboxylate group around the aromatic group, and therefore this cut level is a suitable choice to describe the data set.

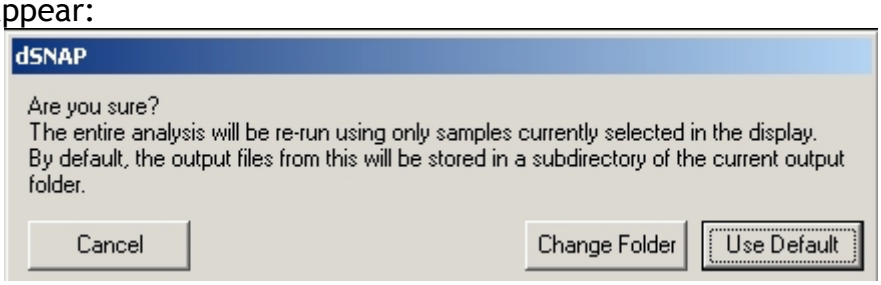
Note:
These geometries are described in the table on the next page.

Cluster	Frequency	Description of geometry
A (red)	19	Dimer interaction involving both nitrogens and both oxygen atoms
B (yellow)	5	Interaction involving only one N and one O – rest of carboxylate group is close to N2 but with no interaction
C (green)	4	Interaction occurs between two planes separated by ~3.2Å
D (cyan)	1	Interaction between N7 and O10, with the planes of the two groups approximately 90° to one another
E (blue)	9	Interaction between N7 and O10 – rest of carboxy group is to the side of C6



However, we know that there is more fine detail. One way we can "drill down" to this detail is to re-run the cluster analysis on selected samples from the data set. In this case we may wish to do this for clusters A (red) and E (blue).

Go to the Cell Display and select cluster E by clicking on its colour in the key on the left. Now go to **Tools > Re-Run Analysis on Selected Samples…** The following dialog box will appear:

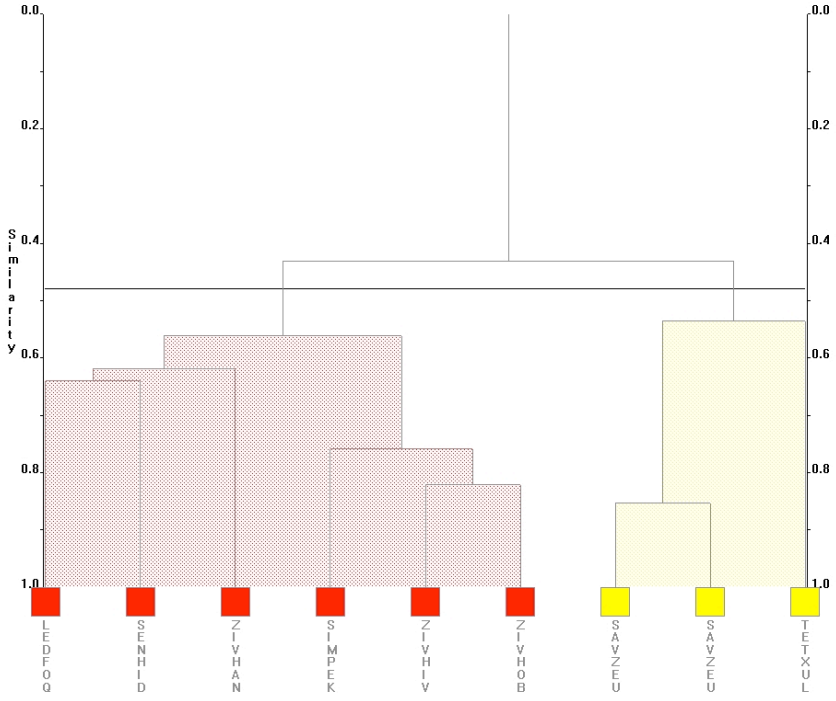


Note:
The **Use Default** option saves the clustering output files in a sub-folder called *Rerun_SelectedSamples* in the output directory for the original analysis.

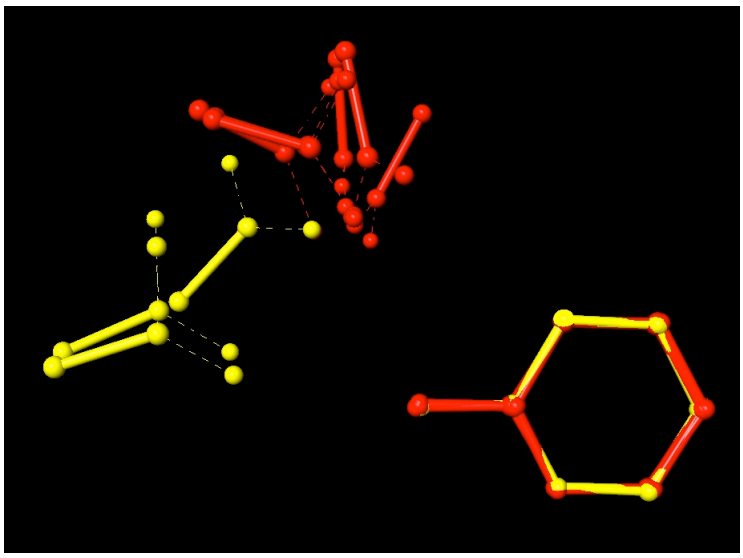
Click on the **Change Folder** button and create a new sub-folder in the output directory. The cluster analysis will be performed as normal and will open on the Cell Display.

Tip:
To avoid confusion, particularly when this process is carried out for several clusters, call the sub-folder for the output by a name which reflects the current cut level of the original dataset, and the cluster being re-clustered, *e.g.* 0688_E.

Look at the dendrogram and 3D plots. The cut level is 0.629, giving rise to 5 clusters which are distinct in space. However, the dendrogram has two main branches. Raising the cut level to a similarity of 0.479 partitions the data into those two groups.



Viewing the groups with the Mulitple Fragments Viewer shows that these groups correspond to the two green striped groups of the original data set at the default cut-level.



To return to the original data set, close the current *d*SNAP window. The *d*SNAP background screen will appear. Clicking on it brings up the Welcome window. Click on the **View Saved Results…** button and navigate to the output folder of the original cluster analysis. Click on the *correlations.txt* file. This will allow you to view the results of the original analysis.

This concludes Example 3.