Interpreting and Validating Results

It is important to remember that *d*SNAP is using a purely geometrical basis to cluster and knows nothing about the chemistry of the dataset, so the user needs to apply his or her chemical knowledge to give real world meaning to the results. The program incorporates many tools that can help the user to interpret the results.

Dendrogram:

There is a seperate quick quide explaining the features of the dendrogram and how it can be used in interpretation. Generally a dendrogram that exhibits clear, distinct clusters suggests a good dataset where there is potential scope to sub-divide the data even further.

3D plot:

A good cluster in the 3D plot should have all of its spheres close together in the plot (see Figure 1.a). Ideally, the cluster should not be diffuse. A 3D plot where a cluster is large and diffuse, or contains spheres which lie relatively far outside the main body of the group may indicate that the cut-level needs to be lowered. A 3D plot where the spheres do not appear to separate into distinct groups suggests that there may not be any structure to the dataset (see Figure 1.b).

Figure 1.a - a 3D Plot displaying good clustering

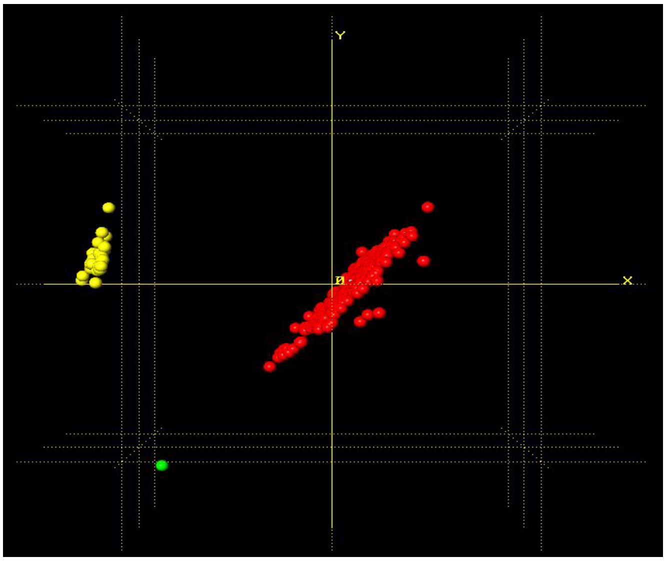
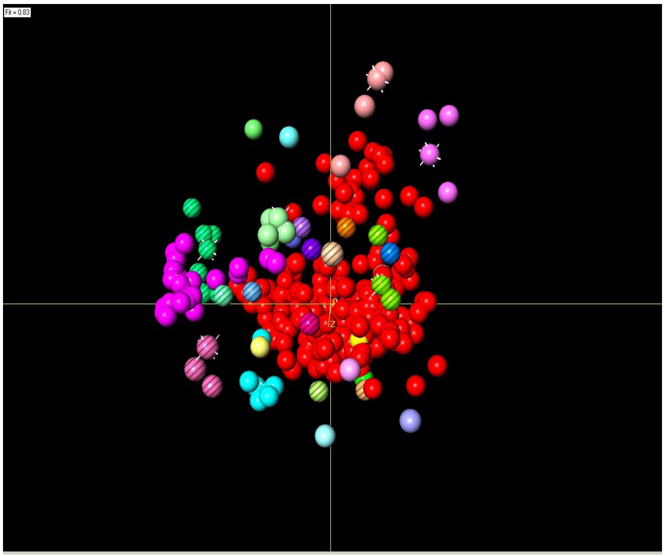


Figure 1.b - a 3D Plot displaying poor clustering



It is useful to switch between the dendrogram and the 3D plot to set the cut-level, as ideally the clustering in the two displays should be consistent with each other *i.e.* seperate clusters should correspond with the groupings observed in the other display.

Visualising structures:

To help assess whether there is chemical sense to the clustering at the chosen cut level, fragments can be viewed using the 3D fragment viewer in *d*SNAP (see seperate quick quide) and an entire individual structure can be viewed in *Mercury*. Looking at the original *ConQuest* search can be useful when the hit fragment is present more than once in a single structure, as it can highlight each hit fragment individually and so can help with identifying fragments.

Variables Space:

In variables space, the numerical results tab can be invaluable in identifying which interatomic distances and angles are particularly influential to the clustering, as clicking on an entry brings up a scatterplot of the correlations between the selected pair of distances and/or angles. Useful results can be obtained for a range of correlation values, not just cases where there is high correlation. It can be useful to pick a pair of parameters that you might expect to be interesting (*e.g.* distances relating to the configuration of a carbon-carbon double bond) as a starting point. The 2D fragment viewer (accessible using the *F2* function key) can be used to highlight which distances or angles a particular parameter refers to.

The tutorial examples work through these methods in a variety of data sets to demonstrate how they can be used to analyse a problem effectively.