Supplementary information for GP5019

Reading Dendrograms and MMDS plots

Dendrograms are useful tools for displaying the results of the clustering calculation analysis using a hierarchical manner of data classification. A dendrogram takes the form of a tree, where each fragment is represented by one of the boxes arranged along the bottom of the plot (see, for example, Figure S1 below). The boxes are joined by horizontal lines, called "tie bars", linking fragments together 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 to be very similar, justifying their being grouped together. If two branches do not meet until near the top of the dendrogram the associated fragments are much less similar and are only loosely related to each other. A set cut-level decides how the dendrogram is split into separate clusters. In this work this cut-level is shown as a solid purple horizontal line (Figure S1). The fragments in a cluster, defined by the cut-level, are arranged with the most similar fragments appearing next to each other and are identically coloured. This representation allows rapid comparison of the different types of fragments and their levels of similarity, both within an individual cluster and within the dataset as a whole.

Metric multidimensional scaling (MMDS) is also used independently of dendrograms to generate a three-dimensional Euclidean space in which each point in this space represents a single fragment. The fragments are then plotted as spheres (see, for example, Figure S1). MMDS preserves the distance metric: fragments whose geometries are very similar lie close to each other, and conversely highly dissimilar fragments are large distances apart. The underlying theory has been described elsewhere (Barr *et al.*, 2005). This assumes, of course, that the dimensionality of the problem can be reduced in this way while still retaining the essential features of the data, and there are checks made for this. To date, this has not been an issue in the range of *d*SNAP studies carried out.

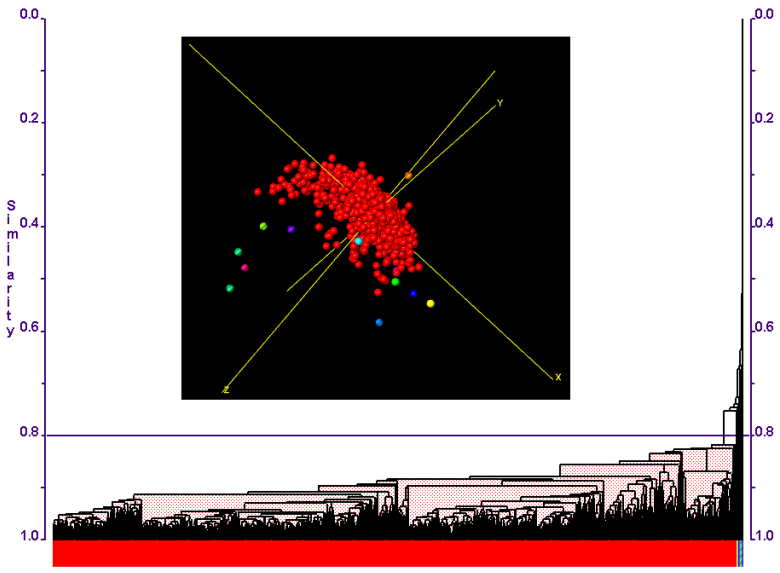

Figure S1: Example dendrogram and MMDS plot (inset) – identical to Figure 3a.

Barr, G., Dong, W., Gilmore, C. J., Parkin, A. & Wilson, C. C. (2005). *J. Appl. Cryst.*, **38**, 833–841.