Acta Crystallographica Section B
Structural Science
ISSN 0108-7681

Using cluster analysis to study transition-metal geometries: four-coordinate complexes with two salicylaldiminato or related ligands

Andrew Parkin,[a]* Gordon Barr,[a] Anna Collins,[a,b] Wei Dong,[a] Christopher J. Gilmore,[a] Peter A. Tasker[b] and Chick C. Wilson[a]

[a]WestCHEM, Department of Chemistry, University of Glasgow, Glasgow G12 8QQ, Scotland, and [b]EaStChem, School of Chemistry, University of Edinburgh, Edinburgh EH9 3JJ, Scotland

Correspondence e-mail: a.parkin@chem.gla.ac.uk

Received 8 January 2007
Accepted 3 May 2007

Cluster analysis is shown to be an effective method to analyse and classify metal coordination geometry in a very large number of four-coordinate *bis*-salicylaldimato (or *bis*-β-iminoketonate) transition-metal complexes available in the Cambridge Structural Database. The methods described require no prior knowledge of chemistry to be input; retrieved structures are automatically clustered into groups based purely on the geometric similarity of the fragments and these groupings can then be interpreted by the structural chemist.

1. Introduction

Despite the ever-increasing number of metal-containing structures within the Cambridge Structural Database (CSD; Allen, 2002), structural analysts fully exploiting the inorganic chemical information available within it are still few and far between. This is, at least in part, because the sheer quantity of information available is somewhat intimidating and extracting the useful facts can be a bewildering and time-consuming task. Thus, the CSD is commonly used as a depository against which it is possible to compare both unit cells and molecular dimensions, allowing the user to highlight unusual values for parameters obtained in structural (or computational) analysis, or to define 'standard' dimensions for a system. These are very important functions, but by the systematic study of related structures and careful examination of the results it is also possible to derive much more meaningful chemical knowledge than that found from such simple searches. Orpen (2002) provides a concise review of such work, with a more recent general review of the developments in inorganic crystal engineering being given by Brammer (2004). Other work in this area includes an excellent study on the coordination of carboxylates by Hocking & Hambley (2005); Fey, Harris *et al.* (2006), Fey, Tsipis *et al.* (2006) and Harris *et al.* (2005) have developed knowledge bases of transition-metal geometries and their associated ligands to add further possibilities for exploiting CSD information. Minguez Espallargas *et al.* (2006) have investigated the interface between inorganic and organic fragments, and have described the intermolecular halogen–halogen contacts in such species; Dance (2003) has correlated observed inorganic intermolecular motifs with their calculated energies. Each of these papers focuses on a few key geometric parameters that are believed to be central to the observed structure, bonding or intermolecular interactions in those particular classes of compounds.

Excellent search and analysis tools are provided by the CSD by the Cambridge Crystallographic Data Centre (CCDC) – these include *ConQuest* (Macrae *et al.*, 2006), *ISOSTAR* (Bruno *et al.*, 1997), Mogul (Bruno *et al.*, 2004), *Vista* (CCDC, 1994) and *MERCURY* (Macrae *et al.*, 2006). Even with these programs, extracting and analysing chemical information from the CSD can be daunting given the volume of data available (59% of Version 5.27 of the CSD entries contain inorganic fragments – 210 942 of a total of 355 064, with 173 227 containing transition metals). Whilst it is usually straightforward to define a search criterion for a particular assembly of connected atoms (a molecular fragment), the subsequent structural analysis of the retrieved data sets (which are often very large) can still present a formidable challenge. As a consequence, searches are frequently carried out with a specific structural correlation in mind and the extracted data interpreted on this basis. This approach is valid, but the question arises as to whether other, less obvious but still relevant, structural correlations may have been overlooked. We show here how the coupling of pattern matching to cluster analysis and multivariate statistical methods described by Barr *et al.* (2005) can be used to classify large numbers of molecular inorganic structures. By applying this method to datasets mined from the CSD, clusters of similar structures are formed on the basis of their underlying geometric properties, which can then be identified and analysed in a chemically meaningful and consistent way.

Cluster analysis and statistical techniques are not new to structural analysis; for example, principal component analysis (PCA), graphical analysis such as scatterplots, and cluster analysis of varying types have all been used to look at a variety of geometric problems, particularly in the study of conformational trends. The use of cluster analysis in this field was first described by Allen & Taylor (1991) and Taylor & Allen (1994); the differences between this pioneering work and the methods applied here have been previously described (Barr *et al.*, 2005). The methods used in this paper have been previously applied to intermolecular interactions (Parkin *et al.*, 2006) and are implemented in the program *dSNAP*, which is available for free download. The initial dataset used with these methods should ideally be as broad as possible; the method then allows both an overview and also an opportunity to 'drill down' into more detailed geometric differences by selecting individual clusters identified in the initial calculation

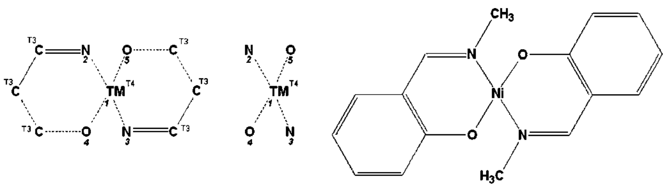

Figure 1
(*a*) The *ConQuest* search fragment defined, illustrating the bond type (a dashed line indicates any bond), and restraints on the coordination number of each atom, with T4 indicating four atoms bound and T3 indicating three atoms bound. The notation used is that of the CCDC. (*b*) The interatomic distances and angles are defined only for the primary coordination sphere. All interatomic distances and angles are calculated in this fragment, not only those that are within conventional bonding distances. Atom numbering within the primary coordination sphere is shown in bold italics beneath each atom. (*c*) A typical structure, bis(*N*-methylsalicylaldiminato)nickel(II).

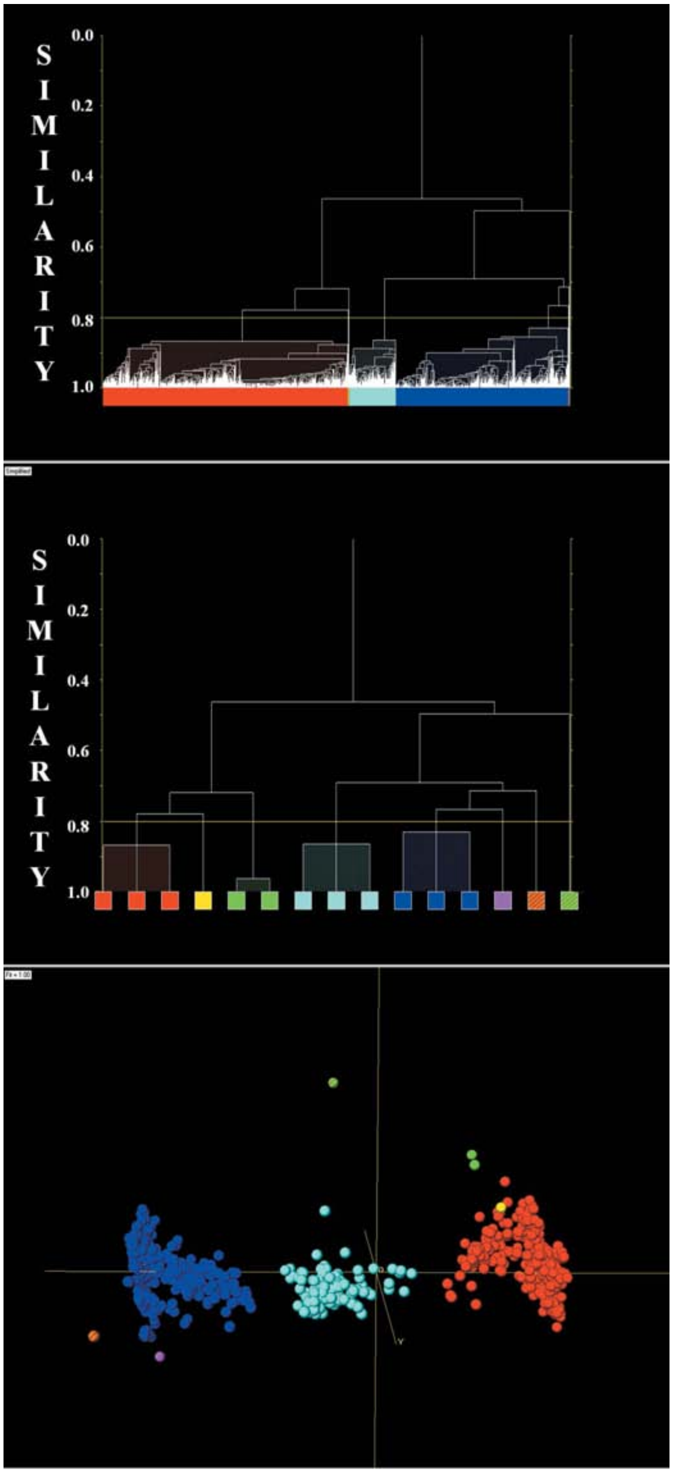

Figure 2
Initial clustering of the dataset with separation at a cut-level of 0.8 in the dendrogram (*a*). The simplified dendrogram (*b*) illustrates the similarity levels at which the different clusters are related and the MMDS plot (*c*) shows well separated and well defined clusters. The colour key to the clusters is the default *dSNAP* colour scheme and is as follows: group *A* red; group *B* yellow; group *C* green; group *D* pale blue; group *E* dark blue; group *F* magenta; group *G* striped orange; group *H* striped pale green.

and then re-clustering these. The observed geometric features can then be studied at an appropriate level of detail to extract the underlying chemical information, and most importantly the groupings suggested will be *free from chemical bias* because no preconceived chemical prejudices have been included in the initial analysis. The methods complement existing techniques and greatly aid the interpretation of data mined from the CSD.

In this paper we present the results of cluster analysis on a large dataset of metal-organic crystal structures. This consists of 1112 fragments from 890 structures, each containing a four-coordinate transition-metal atom with two salicylaldiminato-derived (β-iminoketonate) ligands chelating through the O and the N atoms (Fig. 1). The aim of the analysis is to identify the factors affecting the immediate coordination sphere of the metal, *i.e.* the relative positions of the N and O donor atoms with respect to the metal centre. The majority of the ligands in this study consist of derivatives of that shown in Fig. 1(*c*). Formally the negative charge on the ligand is situated on the coordinating O atom, but there is usually some degree of delocalization of this charge around the ring, leading to a partial ‘averaging’ of the bond lengths in the coordination ring. As the transition metals involved are restricted to being four-coordinate, the principal metal ions present are cobalt, copper, nickel, zinc and palladium, with at least 50 fragments of each metal being present. There are also smaller numbers of fragments containing other metals; manganese (two fragments), iron (two), silver (one), platinum (five) and gold (two). Chemical intuition would lead us to expect three principal geometries – square planar with the N atoms *cis* to one another, tetrahedral, and square planar with the N atoms *trans* to one another. It might be expected that it will be simple to differentiate between transition metals that typically adopt a tetrahedral geometry and those that adopt a square-planar geometry, but some variation within these geometries might also be observed depending on the metal involved.

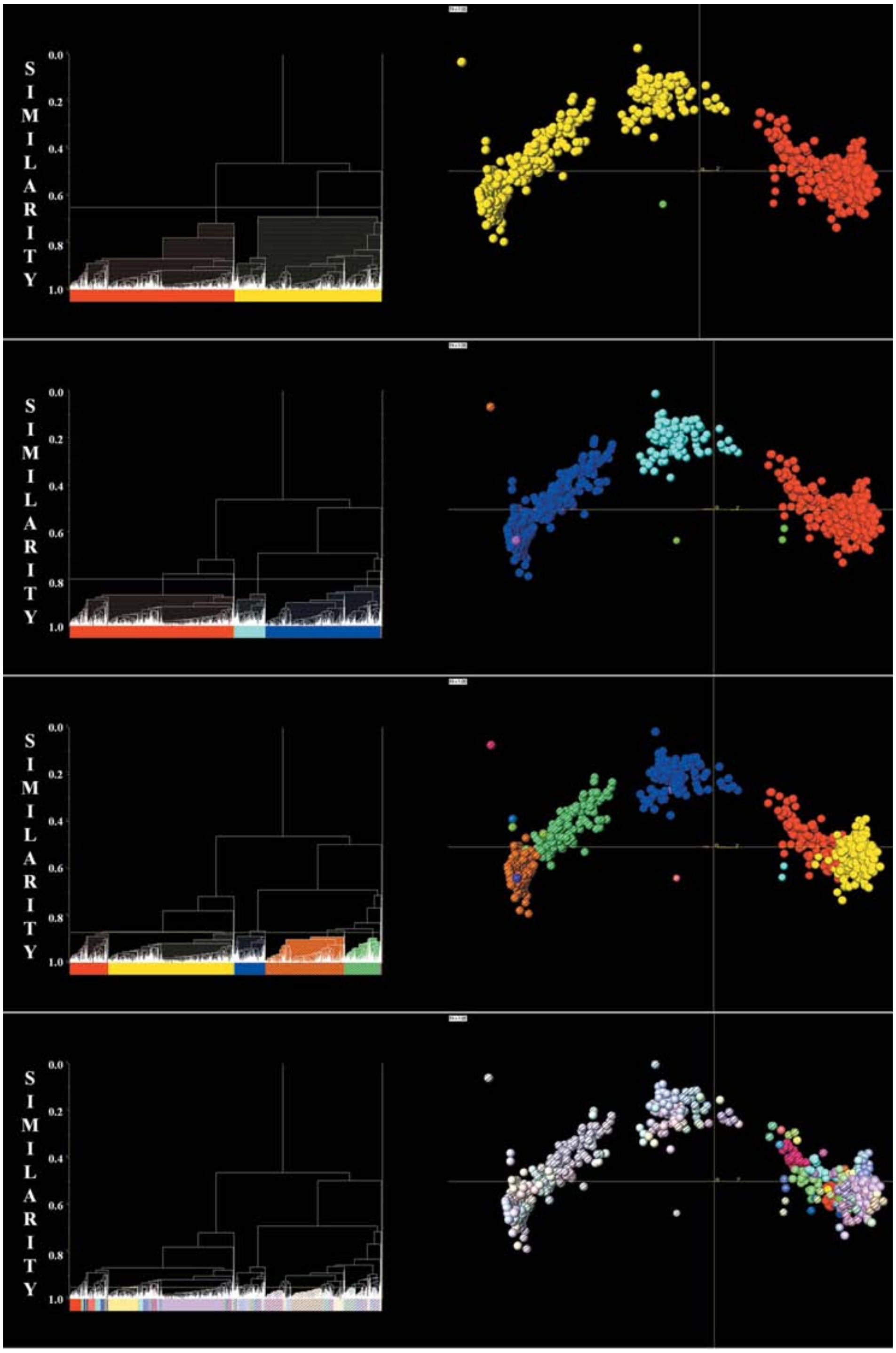

Figure 3
Four different cut-levels are shown on the same dataset at similarity levels of (*a*) 0.65, (*b*) 0.80, (*c*) 0.87 and (*d*) 0.95. The aim is to observe well separated but tight clusters in both the dendrogram (left) and MMDS plot (right). In this instance the best separation is observed in (*b*) at a similarity level of 0.80.

2. Data mining and clustering procedures

2.1. Data mining

The fragment was defined in Version 5.27 of the CSD as a four-coordinate transition-metal atom bound by any bond to four coordinating atoms belonging to two β-iminoketonate-type ligands as in Fig. 1(*a*). The geometric data included in the cluster analysis calculation include all interatomic distances – a total of $(n/2)(n-1)$, where n is the number of atoms – and angles – a total of $(n/2)(n-1)(n-2)$ – for the N_2O_2 primary coordination sphere (Fig. 1*b*). This gives a total of ten interatomic distances and 30 interatomic angles; although this represents a redundancy of geometric information (the fragment is uniquely defined by the ten distances), we have found by experience that differences are accentuated by including additional angle information (Parkin *et al.*, 2006). Although the other atoms of the β-iminoketonate ligands were also defined in the search (Fig. 1*a*), they were not included in the geometric parameter definition (Fig. 1*b*) as our structural interest here resides in the effect of any substitution on the coordination geometry. The only filter applied to the structures is that three-dimensional coordinates should be determined and present in the CSD. A total of 1112 fragments from 890 structures matched these search criteria. Details of the definition of the geometric parameters are available from the *dSNAP* and *ConQuest* files, which are available as supplementary information.[1]

2.2. Dendrogram display and interpretation

Dendrograms are useful tools for displaying the results of the clustering calculation analysis using a hierarchical manner of data classification. It 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, Figs. 2*a* and *b*). 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 cut-level decides how the dendrogram is split into separate clusters. In this work it is shown as a solid, yellow, horizontal line. 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.

The simplified dendrogram quickly gives the analyst an idea of how the groups are related by only including a maximum of three fragments from each cluster, and only showing the similarity levels between the groups. The three fragments used define the extremes of the cluster and the most representative fragment.

2.3. Metric multidimensional scaling

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, Fig. 2*c*). 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 we can reduce the dimensionality of the problem in this way and still retain the essential features of the data, and there are checks made for this. To date, this has not been an issue.

2.4. Methodology of clustering

The clustering presented here was carried out using the *dSNAP* program, which offers a highly visual and interactive way of interrogating geometric data extracted from the CSD.[2] The graphical analysis methods and representations have proven to be both quick and simple to use and highly accurate. Barr *et al.* (2005) have described in greater detail the methods employed and the tools available within the program, and Parkin *et al.* (2006) have previously applied these methods to intermolecular interactions.

In this investigation the dataset was split into a small number of well clustered groups, and chemical similarities within each group were identified that affect the geometry of the metal coordination sphere. The sub-groups were then reclustered and the process undertaken iteratively until all the significant geometric and chemical features had been elucidated. The methods used to estimate the best point at which to cut the dendrogram to define the clusters are described below and some of the important points are discussed in §4; these rely on visibly distinct clusters being observed in both the dendrogram and MMDS plots.

All the similarity levels were chosen to give consistency between the MMDS plot and the dendrogram, *i.e.* well defined clusters in the dendrogram must give equally well defined groups in the MMDS plot. In addition, the fragments were loaded into *ISOSTAR* (Bruno *et al.*, 1997) and coloured by cluster; the resultant plots were checked for chemical sense and consistency. Four examples of the effects on clustering

[1] Supplementary data for this paper are available from the IUCr electronic archives (Reference: BS5042). Services for accessing these data are described at the back of the journal.

[2] These calculations are performed using a program called *dSNAP*. This is a customized version of the computer program *PolySNAP* (Barr *et al.*, 2004*a*,*b*). The software runs on a PC under *Windows*2000© or *XP*©. Although the calculation is elaborate, the total time taken on a 2.4 GHz PC varies between < 1 min for 100 hits and *ca* 1 h for 1500. The current limits are 4000 fragments (n) and 4000 structural parameters (m). If the CSD is installed on the computer that runs *dSNAP*, then structures or groups of structures corresponding to any points on the MMDS plot or the dendrogram can be displayed with the *MERCURY* or *ConQuest* software. The highly visual and interactive nature of the software allows rapid identification of clusters and chemical similarity. The software is available for free download at http://www.bruker-axs.de.

arising from modifying the dendrogram cut-level are shown in Fig. 3. In Fig. 3(*d*) the cut-level is set at too high a similarity level – 127 groups have been formed and it is not possible to separate all of these fully on the corresponding MMDS plot. By contrast, the cut-level shown in Fig. 3(*a*) is also non-ideal: although the samples in the two groups in the MMDS plot are all quite well separated, each of the groups is very diffuse. Fig. 3(*c*) illustrates what appears to be a reasonable cut-level, but the best choice for this dataset is shown in Fig. 3(*b*). In this case the MMDS clusters are both well separated from each other and quite tightly grouped without many structures separated from the cluster centroid. When looking for a cut-level in a dendrogram it is advisable to look initially for a large separation between tie-bars (which signifies a large difference between clusters) – in this case we have chosen the lowest of the large separations – and then compare the clusters in the MMDS plot; in this way it is easy to arrive at an appropriate cut-level such as that shown in Fig. 3(*b*).

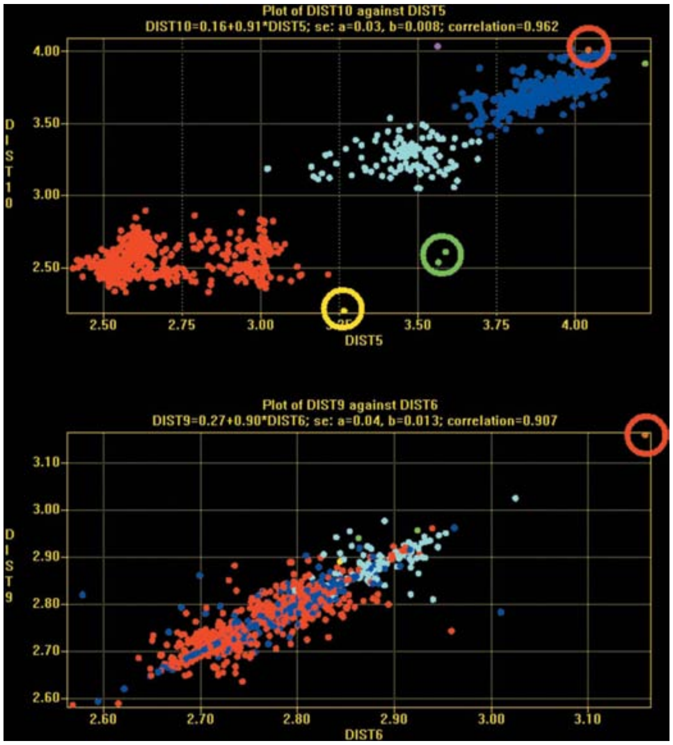

Figure 4
Plots of (*a*) O···O distance (dist10) *versus* N···N distance (dist5) encapsulating the majority of the differences between the different clusters and (*b*) intra-ligand N···O distances (dist6 and dist9) plotted against each other, illustrating how group *G* (circled in red in both plots) is different from the other structures. This figure illustrates why treatment of the entire geometry of the fragment of interest is more useful than relying on a limited number of geometric parameters – it is not immediately obvious from (*a*) that group *G* has a different geometry from the other structures, whereas it is clear within the dendrogram and MMDS plot. Group *B* is circled in yellow and group *C* in green on plot (*a*) to illustrate how their geometries are different from each other and from group *A*. The colours of the individual groups are transferred from the dendrogram in Fig. 2.

The problem of the symmetry of the conformation space of the fragment (also known as topological symmetry or local chemical symmetry) is particularly important in the examples presented (Morgan, 1965; Murray-Rust *et al.*, 1979). An operation was applied to the dataset to bring all fragments into the same volume of conformational space. Every fragment was checked to ensure that the N2···O5 distance (defined as dist7) was shorter than the N3···O4 distance (dist8); if this was not the case N2 and O4 were swapped with N3 and O5, thus ensuring that dist7 was always shorter than dist8. If this is not done then it is possible for identical fragments to appear to lie in different clusters, as they have not been transformed into the same region of symmetry space. Full details of the significance of this issue, and possible methods for its resolution, will be described in a future paper.

3. Results

The initial clustering calculations on all the retrieved fragments gave eight clusters when the dendrogram was cut at the 80% similarity level (Fig. 2). The clusters are well separated and contain well defined groups in the MMDS plot. The three largest correspond to the expected groupings into *cis*-N planar (group *A*), tetrahedral (group *D*) and *trans*-N planar (group *E*) geometries and are well separated in the plot of the N···N distance against the O···O distance in Fig. 4. Other clusters are much smaller containing a maximum of two fragments, and represent more unusual coordination geometries observed in the dataset. Group *B* contains a fragment with *cis*-N planar geometry, but with an unusually short O···O distance and an unusually long N···N distance imposed by the bridge between the imino N atoms. Group *C* also represents two fragments with *cis*-N planar geometry, but while both contain a typical O···O distance, they also contain an unusually long N···N distance, again because of constraints imposed by the ligand. These differences compared with groups *A*, *D* and *E*, and between the two groups, are easily identified in Fig. 4(*a*). Group *F* represents a fragment from a structure with two very different metal environments, despite apparently being similar. In one fragment the *M*—N bond lengths are *ca* 1.8 Å and the *M*—O bond lengths are *ca* 2.0 Å, whereas in the other this trend is reversed. The latter is more probable based on the other structures in the group; despite the low *R* factor reported in the CSD, the former structure is most likely an incorrect entry in the database. Group *G* contains a single fragment with *trans*-N planar geometry with particularly long intra-ligand N···O distances (see Fig. 4*b*). Finally, group *H* represents a silver-containing complex that has a very unusual and irregular coordination geometry, which can only be loosely described as four-coordinate.

A convenient way to represent these differences is in the form of a decision tree, which mimics the dendrogram and describes briefly the principal differences between the clusters. The decision tree in Fig. 5 summarizes the first coordination sphere of the clusters in Fig. 2.

3.1. Sub-clusters; fragments with square-planar *cis*-N geometry

Despite the clear and well separated clusters observed in the initial clustering calculation described above, this does not represent the limits of the classification process. Each of the larger groups can be re-clustered to reveal further sub-clusters, each of which is chemically and structurally distinct. This method of 'drilling down' can be used to access detailed structural information that might be swamped in the previous calculation.

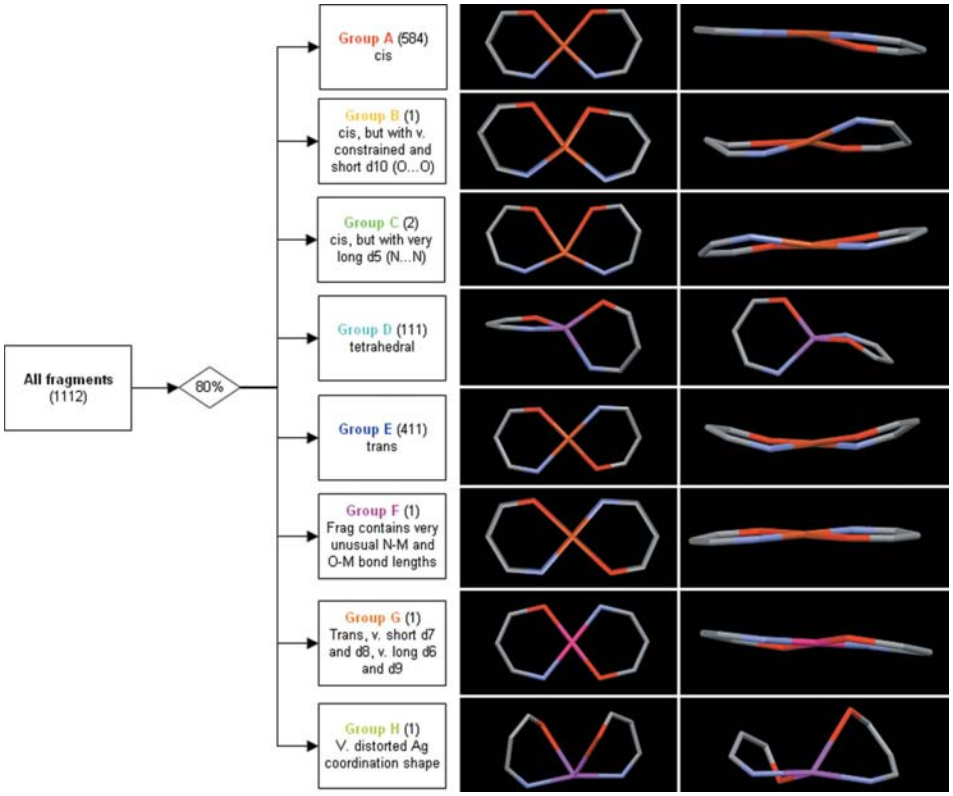
All fragments (1112)
80%
Group A (584) cis
Group B (1) cis, but with v. constrained and short d10 (O...O)
Group C (2) cis, but with very long d5 (N...N)
Group D (111) tetrahedral
Group E (411) trans
Group F (1) Frag contains very unusual N-M and O-M bond lengths
Group G (1) Trans, v. short d7 and d8, v. long d6 and d9
Group H (1) V. distorted Ag coordination shape

Figure 5
Decision tree illustrating the differences between the various clusters in the initial clustering calculation.

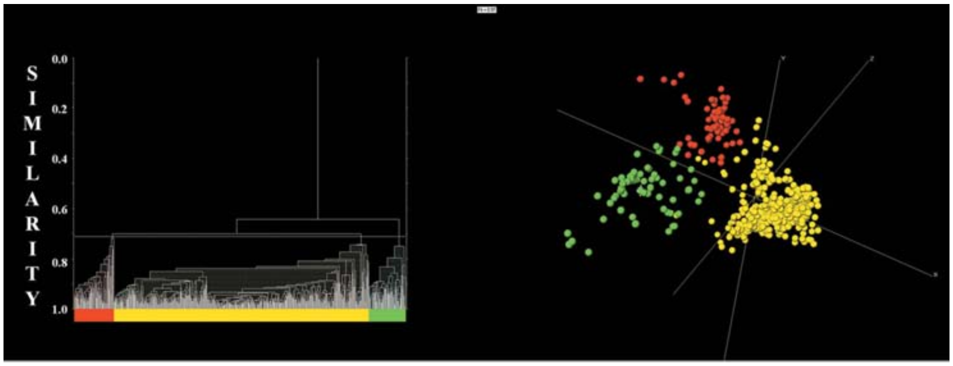

Figure 6
Dendrogram (left) and MMDS plot (right) showing how group *A* can be split into three sub-groups. The colour key to the clusters is the default *dSNAP* colour scheme and is as follows: group *AA* red; group *AB* yellow; group *AC* green.

The 584 fragments making up group *A*, the *cis*-N planar complexes, in the initial calculation form 3 separate sub-groups (Fig. 6) when re-clustered; these sub-groups are most easily summarized in terms of their N···N and O···O distances (Fig. 7). Group *AA* represents those fragments with short O···O and long N···N, group *AC* the fragments with long O···O and long N···N distances and group *AB* those fragments with short N···N distances. The separations between the sub-clusters in the MMDS plot (Fig. 6, right) are less well defined than in the initial clustering, because the differences are less obvious and less distinct than those in the initial calculation.

These 'second-level' distinctions arise from the structural chemistry of the coordinating ligand, illustrated in the decision tree in Fig. 8. The significance of the N···N distance relates principally to the length of the bridge between the N atoms, with a short N···N distance generally signifying a shorter bridge. The fragments in group *AC* have both long O···O and N···N distances because they are forced to have a slight tetrahedral twist; those in group *AA* are still approximately planar, but a longer N···N bridge has forced the O atoms closer together; in the majority of these structures there is also another metal atom being chelated by both O atoms. Of the 70 fragments in group *AA*, 60 have a three-atom aliphatic carbon bridge between the two nitrogen donors. Of the other ten structures, two have a three-atom bridge comprising two aromatic and one aliphatic C atom, three have a bridge of more than three atoms, two are oxime-type N atoms and the OH groups form an intra-complex hydrogen bond, and three have large degrees of steric hindrance from the groups bound to the nitrogen donor atoms. These features require the longer N···N distance. The majority of the fragments in group *AC* have sterically bulky imino substituents displacing the donor atoms towards a tetrahedral arrangement. Group *AB* consists principally of structural fragments with two-atom bridges between the N atoms, with only a few three-atom bridges. Indeed, group *AB* can be re-clustered to account for these finer differences.

3.2. Sub-clusters; fragments with tetrahedral geometry

On re-clustering group *D*, six clusters are observed at the 75% similarity cut-level. The majority of the 111 fragments have approximately perpendicular N—*M*—O coordination planes (groups *DB* and *DC*), with a few fragments adopting a pseudo-tetrahedral arrangement with the geometry being closer to *cis*-N than *trans*-N (group *DA*). Consistent with the other results, these structures are all constrained by large sterically hindering groups or bridges bound to the donor N atoms. There is a larger number of fragments adopting a pseudo-tetrahedral arrangement with their geometry closer to *trans*-N (group *DD*). There are also two fragments identified as outliers, one with very long intra-ligand N···O distances (Group *DE*) and the other with long N—*M* bonds (group *DF*). This information is summarized in Fig. 9.

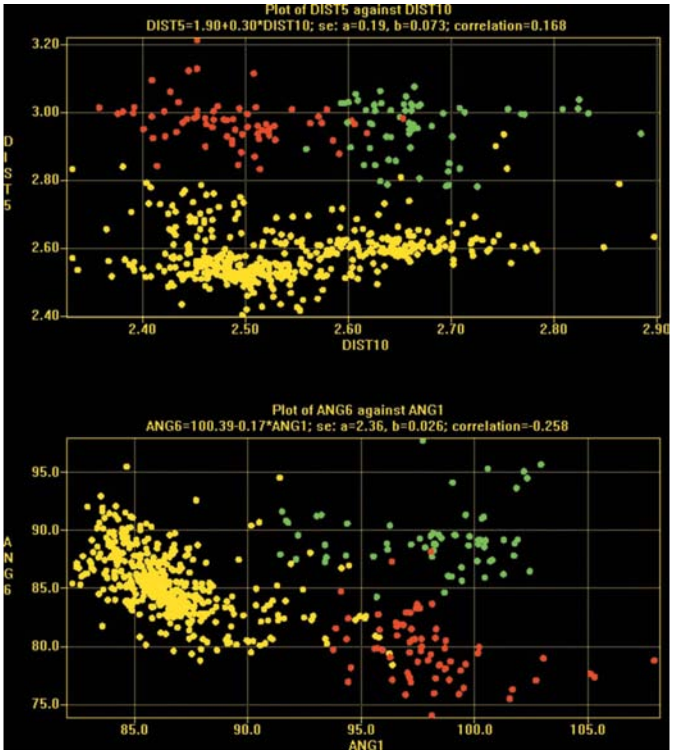

Figure 7
Plots of N···N distance (dist5) *versus* O···O distance (dist10) (graph *a*, top) and O—*M*—O (ang6) *versus* N—*M*—N (ang1) (graph *b*, bottom). The colours used are transferred from the dendrogram in Fig. 6.

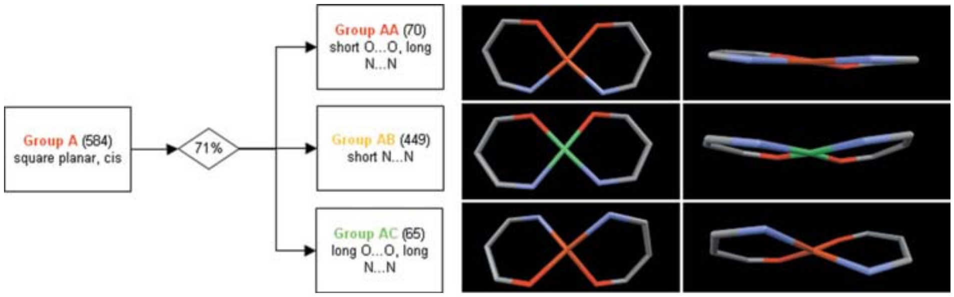

Figure 8
Decision tree illustrating the differences between the various clusters in group *A*.

3.3. Sub-clusters; fragments with square-planar *trans* geometry

On re-clustering group *E*, five clusters are observed at the 73% similarity cut-level. The two larger clusters contain a planar donor set (group *EA*, 279 fragments) and a planar donor set with a significant distortion towards tetrahedral geometry (group *EE*, 128 fragments). The other three groups are outliers: group *EB* represents a fragment with particularly long intra-ligand N···O distances; both fragments in group *EC* originate from a highly disordered structure; group *ED* represents a fragment with one particularly long intra-ligand N···O distance, and upon consideration of the local geometry around the N atom, appears to contain a coordinate error. This information is summarized in Fig. 10.

4. Choosing a suitable cut-level, and other cautionary notes

Estimating the number of clusters reliably is an unsolved problem in classification methods. As with all clustering and classifying methods, there is a danger of overclassifying the dataset under investigation and, taken to the extreme, cluster analysis can be used to show that each sample in the dataset forms its own unique cluster. At the other extreme, even very different fragments can be included in a single cluster, simply by virtue of being in the same dataset. Although neither of these extremes is actually incorrect, in most cases they are unlikely to be effective in describing the dataset (or in extracting important structural chemical information), and it is more realistic to choose a clustering level somewhere in-between. The question therefore arises: how do we choose an appropriate level of clustering? The answer lies in how the observed clusters at the chosen level are interpreted.

There are two possible philosophies that can be adopted. The first involves attempting to determine all differences between all groups in the analysis in a single clustering calculation. However, it is our experience that the most effective clustering philosophy is to cluster the dataset initially into a small number of distinctly different groups, and then by repeating the calculation on subsets it often becomes possible to investigate more subtle, but still significant, differences. Small differences risk being deemed insignificant using the first method, but are easily observed using the second.

The quality of the results obtained is dependent on two major factors: the quality of the input data and the quality of the analysis. In terms of input data quality, it is important to inspect unusual motifs visually before deciding whether to keep them in the analysis. The ability to interpret these structures using chemical knowledge is necessary to distinguish between unusual and possibly interesting structures, and poorly or incorrectly determined structures. These errors (as opposed to true outliers) will usually manifest themselves in terms of some highly unlikely geometric data. Consequently, it is important to be careful when including H atoms in analysed geometric parameters as they are often determined with poor precision. Often the best course of action is to include a H atom in the search fragment, but only define geometric parameters for analysis of the heavier atoms. The quality of the analysis depends on an awareness of chemical structure and bonding in determining the relevance of the results. Here we have included some detail on the sub-clustering of the principal groups, but in some cases there may be little to gain by doing this; equally in other cases it may be possible to drill down further to extract more detailed chemical information from the system. It is helpful at the outset to decide the goal of the analysis, and if there is no obvious additional structural information to be obtained from further detailed clustering then perhaps the analysis has been completed. Being aware of these factors should be enough to prevent the overanalysing of a system.

5. Conclusions

Database mining and the application of *dSNAP* has made it relatively simple to identify the three principle geometries of *bis*(β-iminoketonato)metal complexes. The adoption of these structures and more subtle variations of them is very dependent on constraints imposed by the ligands, in particular the presence or absence of a two- or three-atom bridge between two imino N atoms leads to *cis*-N square planar or *trans*-N square planar structures, respectively. Deviations from planarity depend upon both the nature of the metal ion and the presence of steric constraints imposed by the ligands. Other differences are represented by sub-clusters of the three major structural types. Several of these differ in the relative sizes of the N$\cdots$N and O$\cdots$O non-bonded distances in the coordination spheres and again a combination of effects imposed by variations of ligand superstructures (the nature of linkages between the imino donor atoms and whether imino nitrogen substituents impart steric hindrance or attraction between the chelating units), and the coordination preferences of the complexed metals can account for most of these variations. We are currently exploring other methods of analysing this type of data in *dSNAP*.

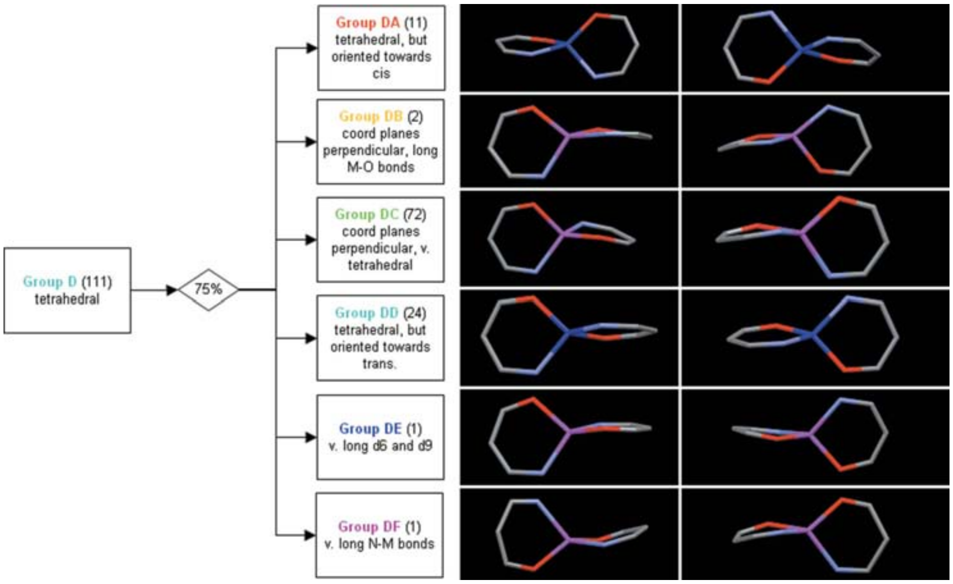
Group D (111)
tetrahedral
75%
Group DA (11)
tetrahedral, but oriented towards cis
Group DB (2)
coord planes perpendicular, long M-O bonds
Group DC (72)
coord planes perpendicular, v. tetrahedral
Group DD (24)
tetrahedral, but oriented towards trans.
Group DE (1)
v. long d6 and d9
Group DF (1)
v. long N-M bonds

Figure 9
Decision tree for the tetrahedral structures, illustrating the differences between the various clusters.

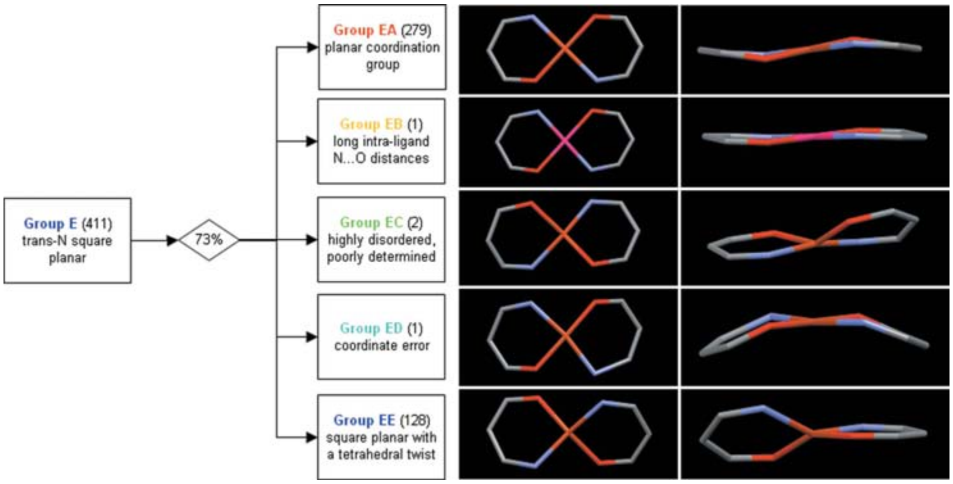
Group E (411)
trans-N square planar
73%
Group EA (279)
planar coordination group
Group EB (1)
long intra-ligand N...O distances
Group EC (2)
highly disordered, poorly determined
Group ED (1)
coordinate error
Group EE (128)
square planar with a tetrahedral twist

Figure 10
Decision tree illustrating the differences between the various clusters in group *E*.

We have shown that cluster analysis can be used quickly and efficiently to classify interactions between a metal and the coordinating ligands, allowing significant structural chemical information to be extracted without prior knowledge of the coordination chemistry of a given ligand type. Large and complex datasets can be handled simply because the method interprets the structure as a whole, rather than looking at individual parameters. The facility to also look at small clusters allows the fast identification of unusual structural features and defines structures that contain errors in their structure determination or documentation. The analysis requires no chemical input and interprets the structural fragments solely on the basis of their relationship to others in the set, so that no chemical bias is introduced at the clustering stage. The method works best by starting with a 'broad-brush' approach, including more rather than fewer structures, and then 'drilling down' to find the detailed differences between individual clusters. The method is greatly aided by the highly visual and interactive nature of the displays in the *dSNAP* program. All *dSNAP* and CSD search files used in this paper are available as supplementary information.

References

Allen, F. H. (2002). *Acta Cryst.* B**58**, 380–388.

Allen, F. H. & Taylor, R. (1991). *Acta Cryst.* B**47**, 404–412.

Barr, G., Dong, W. & Gilmore, C. J. (2004*a*). *J. Appl. Cryst.* **37**, 658–664.

Barr, G., Dong, W. & Gilmore, C. J. (2004*b*). *J. Appl. Cryst.* **37**, 874–882.

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

Brammer, L. (2004). *Chem. Soc. Rev.* **33**, 476–489.

Bruno, I. J., Cole, J. C., Lommerse, J. P. M., Rowland, R., Taylor, R. & Verdonk, M. L. (1997). *J. Comput. Aided Mol. Des.* **11**, 525–537.

Bruno, I. J., Cole, J. C., Kessler, M., Luo, J., Motherwell, W. D. S., Purkis, L. H., Smith, B. R., Taylor, R., Cooper, R. I., Harris, S. E. & Orpen, A. G. (2004). *J. Chem. Inf. Comput. Sci.* **44**, 2133–2144.

CCDC (1994). *Vista. A Program for Analysis and Display of Data Retrieved for the CSD*. Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge, England.

Dance, I. (2003). *CrystEngComm*, **5**, 208–221.

Fey, N., Harris, S. E., Harvey, J. N. & Orpen, A. G. (2006). *J. Chem. Inf. Model.* **46**, 912–929.

Fey, N., Tsipis, A. C., Harris, S. E., Harvey, J. N., Orpen, A. G. & Mansson, R. A. (2006). *Chem. Eur. J.* **12**, 291–302.

Harris, S. E., Orpen, A. G., Bruno, I. J. & Taylor, R. (2005). *J. Chem. Inf. Model.* **45**, 1727–1748.

Hocking, R. K. & Hambley, T. W. (2005). *J. Chem. Soc. Dalton Trans.* **5**, 969–978.

Macrae, C. F., Edgington, P. R., McCabe, P., Pidcock, E., Shields, G. P., Taylor, R., Towler, M. & de Streek, J. (2006). *J. Appl. Cryst.* **39**, 453–457.

Minguez Espallargas, G., Brammer, L. & Sherwood, P. (2006). *Angew. Chem. Int. Ed.* pp. 435–440.

Morgan, H. L. (1965). *J. Chem. Doc.* **5**, 107–113.

Murray-Rust, P., Burgi, H.-B. & Dunitz, J. D. (1979). *Acta Cryst.* A**35**, 703–713.

Orpen, A. G. (2002). *Acta Cryst.* B**58**, 398–406.

Parkin, A., Barr, G., Dong, W., Gilmore, C. J. & Wilson, C. C. (2006). *CrystEngComm*, **8**, 257–264.

Taylor, R. & Allen, F. (1994). *Structure Correlation*, edited by H.-B. Burgi & J. D. Dunitz, Vol. 1, pp. 111–161. Weinheim: VCH.