Acta Crystallographica Section B
Structural Science
ISSN 0108-7681

General method for the description, visualization and comparison of metal coordination spheres: geometrical preferences, deformations and interconversion pathways

Jing Wen Yao,[a,b] Royston C. B. Copley,[a]† Judith A. K. Howard,[a] Frank H. Allen[b]* and W. D. Samuel Motherwell[b]

[a]Department of Chemistry, University of Durham, South Road, Durham DH1 3LE, England, and [b]Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, England

† Present address: Computational, Analytical and Structural Sciences, GlaxoSmithKline Pharmaceuticals, New Frontiers Science Park (North), Third Avenue, Harlow, Essex CM19 5AW, England.

Correspondence e-mail: allen@ccdc.cam.ac.uk

Received 16 November 2000
Accepted 6 March 2001

The coordination sphere geometry of metal atoms (M) in their complexes with organic and inorganic ligands (L) is often compared with the geometry of archetypal forms for the appropriate coordination number, n in ML_n species, by use of the $k = n(n-1)/2$ $L-M-L$ valence angles subtended at the metal centre. Here, a Euclidean dissimilarity metric, $R_c(x)$, is introduced as a one-dimensional comparator of these k-dimensional valence-angle spaces. The computational procedure for $R_c(x)$, where x is an appropriate archetypal form (*e.g.* an octahedron in ML_6 species), takes account of the atomic permutational symmetry inherent in ML_n systems when no distinction is made between the individual ligand atoms. It is this permutational symmetry, of order $n!$, that precludes the routine application of multivariate analytical techniques, such as principal component analysis (PCA), to valence angle data for all but the lowest metal coordination numbers. It is shown that histograms of $R_c(x)$ values and, particularly, scatterplots of $R_c(x)$ values computed with respect to two or more different appropriate archetypal forms (*e.g.* tetrahedral and square-planar four-coordinations), provide information-rich visualizations of the observed geometrical preferences of metal coordination spheres retrieved from, *e.g.* the Cambridge Structural Database. These mappings reveal the highly populated clusters of similar geometries, together with the pathways that map their geometrical interconversions. Application of $R_c(x)$ analysis to the geometry of four- and seven-coordination spheres provides information that is at least comparable to, and in some cases is more complete than, that obtained by PCA.

1. Introduction

Systematic knowledge about the geometries of metal coordination polyhedra ML_n is of fundamental importance in inorganic chemistry (see *e.g.* Drew, 1977; Kepert, 1987). Such data permit general classifications, analysis of deformations from archetypal forms and the study of interconversion pathways that connect these forms. Given the vast amount of crystallographic data for complexes of transition metals that now exist in the Cambridge Structural Database (CSD: Allen, Davies *et al.*, 1991; Allen & Kennard, 1993), a major problem is to devise appropriate methodologies for describing and, in particular, for mapping and visualizing the multi-dimensional geometrical characteristics of subsets of this data.

Conceptually, these issues are analogous to the study of conformational preferences, deformations and interconversion pathways in flexible organic systems. Conformational problems are now studied routinely using structure correlation techniques (see *e.g.* Bürgi & Dunitz, 1994) applied to data

retrieved from the CSD. Applications to both acyclic and cyclic systems involve the analysis of torsional datasets (N_t torsional descriptors for each of the N_f examples of the chemical system in the CSD). For low-dimensional problems, scattergrams of pairs of torsion angles provide mappings of conformational space, but for $N_t = 5$ torsions multivariate statistical methods such as principal component analysis (PCA: Chatfield & Collins, 1980; Allen *et al.*, 1991*a*,*b*,*c*) are crucial in reducing the dimensionality of the problem.

A number of geometrical constructs and metrics have been proposed to describe metal coordination spheres, usually involving comparisons with the geometries of ideal regular archetypes (Lueken *et al.*, 1987; Pinsky & Avnir, 1998). Some of these measures apply only to a given archetype (x), *e.g.* an octahedron, while others apply to a given coordination number n. Recently, more general measures have been proposed, such as the polyhedral sphericity measures of Balic Zunic & Makovicky (1996), later extended to describe the regularity of the distribution of ligand atoms, irrespective of the position of the central metal atom in the polyhedron (Makovicky & Balic Zunic, 1998). While these specialized metrics are valuable, they are relatively complex to compute and apply in systematic studies of large numbers of structures retrieved *e.g.* from the CSD.

For ML_n coordination, considerable knowledge is implicit in the $N_a = n(n-1)/2$ values of the $L—M—L$ valence angles. Just as torsion angles are the natural descriptors of chain and ring conformations, the $L—M—L$ valence angles are the natural descriptors of shape for metal coordination spheres, and are used by many structural chemists for that purpose. They are simple to calculate and tabulate for systematic analyses using crystallographic databases, and their value has been demonstrated (Taylor & Allen, 1994; Klebe & Weber, 1994; Auf der Heyde & Bürgi, 1989) for the lower coordination numbers, n = 3, 4, 5, using statistical and numerical methods analogous to those applied to the conformational analysis of organic systems.

However, one of the main problems in applying multivariate methods to structural data arises from the topological symmetry of the chemical substructure of interest (Allen, Doyle & Taylor, 1991*a*,*b*,*c*; Taylor & Allen, 1994). This gives rise to increasing numbers of possible enumerations for the atoms of the substructure as n (ring size or ligand count) increases, and for carbocyclic rings of size n, there are $2n$ possible atomic enumerations which must be considered in comparing one conformation with another. For ML_n systems, $n!$ atomic enumerations must be considered in comparing a given experimental result with the geometry of a fixed archetype defined in terms of a fixed enumeration of the ligand atoms. Thus, the visualization and analysis of ML_n angular data becomes increasingly intractable, as summarized in §2.

In this paper, we present a general computational procedure for the analysis of angular data for ML_n systems, which takes account of topological symmetry. Comparison of individual coordination spheres with those of a fixed archetype (x) is achieved by the use of a symmetry-modified Euclidean dissimilarity metric, $R_c(x)$, based on the $L—M—L$ valence angles. The method represents a major modification of dissimilarity techniques proposed earlier (Allen *et al.*, 1994; Copley, 1995). $R_c(x)$ values can be used alone to visualize the geometrical complexity of an ML_n system or can form the basis for the systematic application of other techniques, such as PCA. The method has been described briefly elsewhere and proved by a preliminary application to ML_7 systems (Howard *et al.*, 1998). This paper provides a complete description of the computational procedures and illustrates their application to ML_4 systems, normally tractable to standard techniques, and expands the discussion of ML_7 systems, in which topological symmetry presents a major problem. An analysis of ML_3 systems is in preparation.

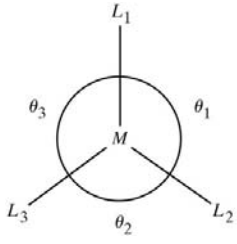
2. Methodology

2.1. Archetypal geometries for ML_n systems (n = 3–7)

Archetypal forms of metal coordination spheres are described by standard angles θ_{sk}, the $k = n(n-1)/2$ valence angles subtended at the metal centre by pairs of ligands, L, in an ML_n system. In many cases θ_{sk} are fixed by the symmetry of the archetype, *e.g.* trigonal planar (TP, D_{3h}, $n = 3$, $k = 3$, $\theta_{sk} = 120°$), tetrahedral (T, T_d, $n = 4$, $k = 6$, $\theta_{sk} = 109.5°$), square planar (SQP, D_{4h}, $n = 4$, $k = 6$, $\theta_{sk} = 90, 180°$), square pyramidal (SQPy, C_{4v}, $n = 5$, $k = 10$, $\theta_{sk} = 90, 180°$), trigonal bipyramidal (TBP, D_{3h}, $n = 5$, $k = 10$, $\theta_{sk} = 90, 120°$), octahedral (O, O_h, $n = 6$, $k = 15$, $\theta_{sk} = 90, 180°$). For seven-coordination, the archetypal forms commonly used are the pentagonal bipyramid (PBP, D_{5h}, $n = 7$, $k = 21$, $\theta_{sk} = 72, 90, 144, 180°$), the capped octahedron (COC, C_{3v}) and the capped trigonal prism (CTP, C_{2v}). In these latter two cases, and indeed for some other coordination numbers, the symmetry does not fix the angles of the archetypal forms. The choice of standard reference angles in the CTP case is described in a later section, while more general comments on the choice of reference angles are given in §5.

2.2. Topological symmetry and PCA in ML_n systems

The ML_3 substructure of Fig. 1 is topologically symmetric and the topological equivalence of the ligand atoms L means that the $n! = 6$ atomic numbering schemes shown are also equivalent. As it is impossible to impose a unique atomic numbering when performing a substructure search of a chemical database such as the CSD, then it is not possible to



Valence angle permutations in ML_3 systems

n	Permutation		
1	θ_1	θ_2	θ_3
2	θ_1	θ_3	θ_2
3	θ_2	θ_1	θ_3
4	θ_2	θ_3	θ_1
5	θ_3	θ_1	θ_2
6	θ_3	θ_2	θ_1

Figure 1
Topological symmetry in ML_3 systems.

ensure that the intraligand valence angles computed from a given database hit are generated in a consistent order. In this example, six possible valence-angle orderings are possible and the search process will choose one of these orderings randomly for each hit, *i.e.* the search process will assign each hit to one of the six possible asymmetric units of the valence-angle parameter space. This causes difficulties in applying multivariate numerical analysis techniques to the raw output from a CSD search, and a general method for obviating these difficulties in *e.g.* principal component analysis (PCA: Chatfield & Collins, 1980; Allen, Doyle & Taylor, 1991*a*,*b*,*c*) is to use the atomic permutational symmetry to fill all of the asymmetric units of parameter space, *i.e.* to perform a sixfold expansion of the resulting matrix of valence angles. Such expansions are commonly used in the PCA of torsional data for *n*-membered carbocyclic rings, where the size of the data matrix is limited to a 2*n*-fold expansion of the raw data from the search process. In applying PCA to the valence-angle spaces of metal coordination sphers, *n*!-fold expansions of the basic data are clearly required. These expansions involve *n*! = 120, 720, 5040 and 40 320 permutational isomers for *n* = 5, 6, 7, 8 coordination, yielding data matrices that rapidly become impossible to handle routinely in many data analysis packages, including the CSD system program *Vista* (Cambridge Structural Database, 1995).

2.3. The angular discrepancy index $R_c(x)$

Similarity and dissimilarity metrics are commonly used in comparing chemical systems (Johnson & Maggiora, 1990) and form the basis for numerical techniques such as multivariate cluster analysis (Everitt, 1980; Taylor & Allen, 1994). In order to compare individual observations of coordination sphere geometry from crystal structures with standard geometries, we use the Euclidean distance metric $R_c(x)$, expressed as a percentage value using the following expression

$$R_c(x) = 100 \min\left\{\frac{\sum[\theta_{ok} - \theta_{sk}]^2}{\sum \theta_{sk}^2}\right\}_p^{1/2}, \qquad (1)$$

in which (x) identifies the archetypal comparator (*e.g.* T, SQP, TBP *etc.*), the θ_{sk} are the standard intra-ligand angles for this archetype, the θ_{ok} are the observed values from a specific crystal structure, and $k = 1 \rightarrow n(n-1)/2$ for an ML_n centre.

In order to accommodate topological symmetry, we have written a local *FORTRAN*77 code to generate all the p permutations of the valence angle sequence that correspond to the n! possible atomic enumerations of the ML_n substructure. The $R_c(x)$ index is computed using fixed archetypal angles θ_{sk} for each permutation of the observed valence angles θ_{ok}. The minimum value of $R_c(x)$ over all permutations quantifies the dissimilarity of the observed experimental crystal structure geometry from that of the chosen reference archetype and, by default, the atomic numbering that gives rise to this minimum is stored, since it represents the unique numbering for that substructure, *i.e.* the atomic numbering which, when imposed onto the corresponding crystal structure, yields the angular sequence which is closest to that of the archetype. For a number of ML_n coordinations, there exist two or more standard archetypes, and $R_c(x)$ can be computed against each reference geometry, *e.g.* $R_c(T)$ and R_c(SQP) for ML_4 systems, and used together in systematic analyses.

2.4. Database search and retrieval

Structural data for transition metal coordination complexes were retrieved from the CSD (various versions from 1996 to 2000) using the program *Quest*3D (Cambridge Structural Database, 1994). Substructure searches were constructed using a specific central metal atom, or the group element symbol TR (any transition metal), connected to the required number of ligand atoms, L, defined as any atom except hydrogen or a transition metal. A 'total coordination number' setting was used to retrieve crystal structure observations having exactly the required coordination. Substructure searches were further constrained using secondary search criteria so that hits (*a*) had error-free coordinate sets from CSD validation procedures, (*b*) were not disordered, (*c*) had a perfect match between their chemical and crystallographic connectivity representations, and (*d*) had a crystallographic $R < 0.10$. For each hit, the *Quest*3D output comprised (*a*) the $L-M-L$ valence angles, (*b*) the $M-L$ distances and (*c*) the atomic coordinates for the further calculations required by the permutational $R_c(x)$ generation procedure.

3. Application to ML_4 systems

3.1. Dataset and PCA results

The dataset used here comprised 127 randomly chosen CuL_4 complexes in which L is any atom except H and all ligands are unidentate. More complete analyses of CuL_4 datasets have been carried out by Klebe & Weber (1994) and Raithby *et al.* (2000), and the dataset used here was chosen for example purposes only. PCA was applied to a data matrix that had been fully expanded using permutational symmetry and

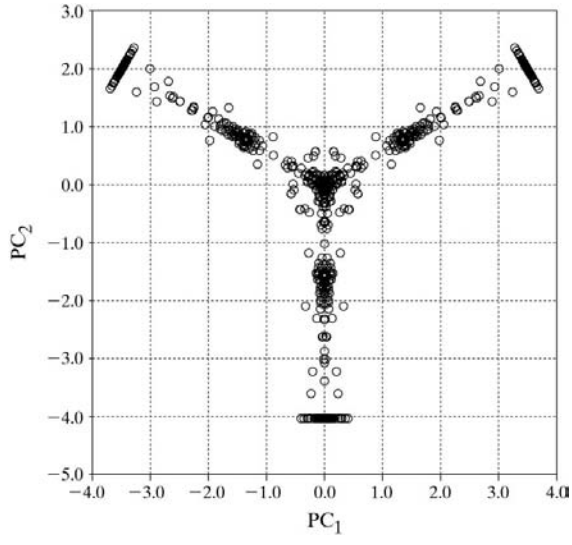


Figure 2
PC1 *versus* PC2 for test set of 127 CuL_4 complexes.

Table 1
Results of the principal component analysis of the test set of 127 CuL_4 complexes.

Cumul. var. is the cumulative variance.

PC's	Eigenvalue	Variance (%)	Cumul. var. (%)
1	2.832	47.2	47.2
2	2.832	47.2	94.4
3	0.209	3.49	97.9
4	0.042	0.70	98.6
5	0.042	0.70	99.3
6	0.042	0.70	100.0

comprising the six $L-\mathrm{Cu}-L$ angles for 127 × 24 = 3048 observations that fill valence angle space. PCA was carried out using the CSD System program *Vista* (Cambridge Structural Database, 1994) and results are summarized in Table 1. The two degenerate PCs, PC1 and PC2, together account for 94.4% of the variance in the dataset, with PC3 accounting for a further 3.5%. The PC1 *versus* PC2 mapping illustrated in Fig. 2 represents a view along the threefold axis of a tetrahedron. The central cluster arises from tetrahedral (T) complexes, while the other clusters of note are the trigonally symmetric set that arise from square planar (SQP) examples. Reference back to the CSD shows that Cu^{I} species account for the T cluster, while Cu^{II} species account for the SQP cluster (Raithby *et al.*, 2000). Most importantly though, the PC map shows two other significant features:

(i) a number of experimental observations fall on lines that connect the expected archetypal T and SQP forms, mapping the geometrical changes that accompany interconversion between the two forms, and

(ii) the SQP clusters reveal a range of distortions along an axis perpendicular to the interconversion pathway, while the T cluster is represented by a compact circle of data points.

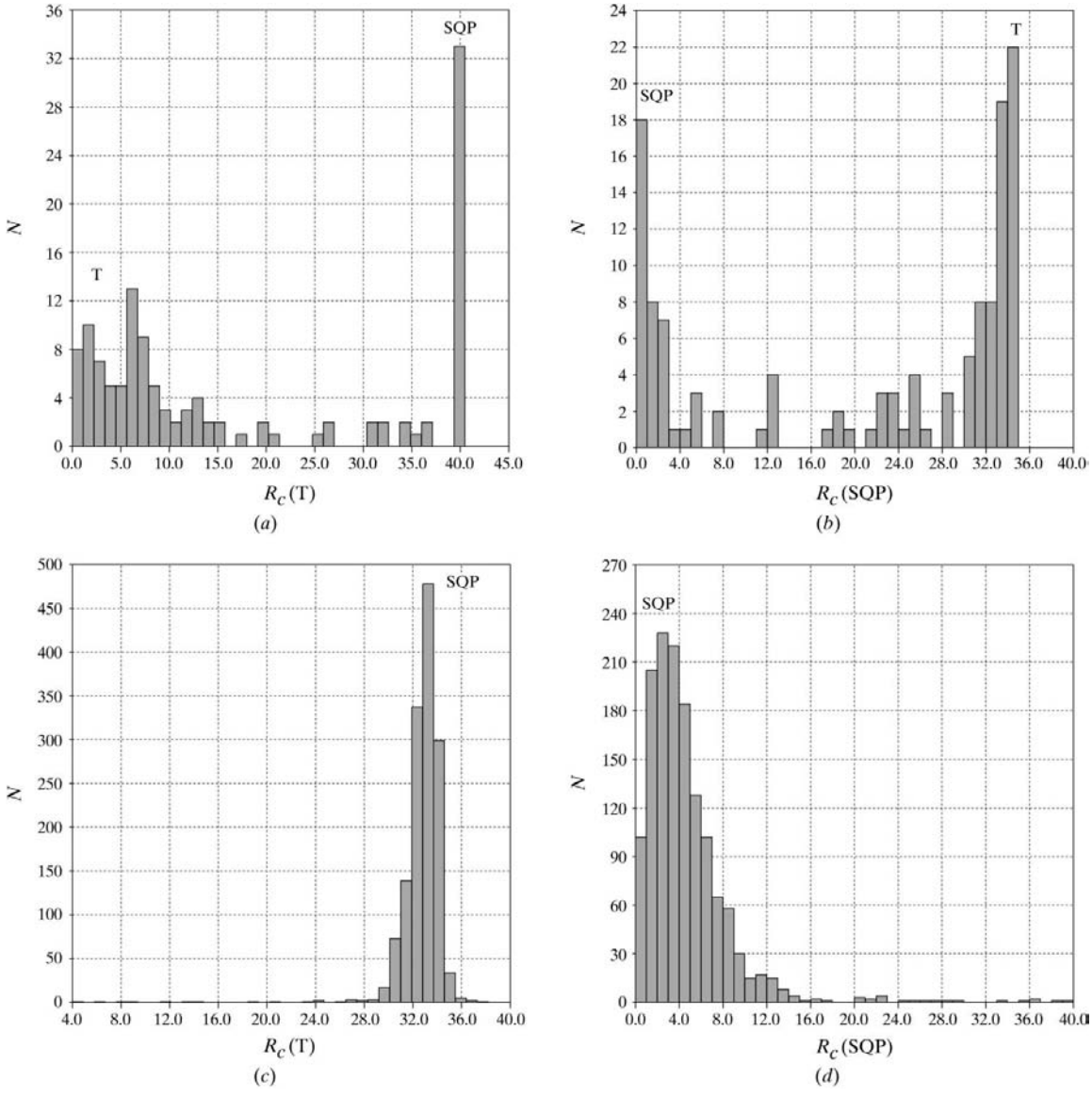


Figure 3
(*a*) R_c(T) and (*b*) R_c(SQP) histograms for 127 CuL_4 complexes, and (*c*) R_c(T) and (*d*) R_c(SQP) for 1406 PtL_4 complexes.

These SQP distortions can be traced to complexes having bite angles that are much smaller/larger than the standard of 90°, due to the formation of three-membered metallacycles.

Application of 24-fold symmetry-expanded PCA to 127 CuL_4 fragments poses no real computational problems and one PC mapping provides an information-rich visualization of the geometrical diversity in the dataset. However, merely increasing the size of the sample to 417 experimental examples now generates a symmetry-expanded six-dimensional data matrix that must encompass the 24 × 417 = 10 008 observations that fill the valence-angle space. This relatively small basic dataset exceeds the current capabilities of the *Vista* program and we must examine other methods of visualizing structural diversity.

3.2. $R_c(x)$ analysis

While we envisage $R_c(x)$ analysis being used for larger datasets, we have computed R_c(T) and R_c(SQP) for the 127 examples of the test CuL_4 dataset, minimized over permutational symmetry as in (1), so as to compare and contrast the value of $R_c(x)$ analysis with PCA. In Fig. 3 we show histograms of (*a*) R_c(T) and (*b*) R_c(SQP). Both show two clear peaks, corresponding to the T and SQP examples, with a small number of observations which connect these peaks. We contrast this behaviour with that of 1406 PtL_4 complexes from the CSD (Figs. 3*c* and *d*), where the predominance of SQP spheres at the d^8 Pt^{II} centre is clearly illustrated. Thus, the structural diversity of the samples is revealed at a gross level, irrespective of which archetype of four-coordination is used as the standard.

In our earlier attempts (Allen *et al.*, 1994) to detect conformational diversity among cyclic and acyclic substructures retrieved from the CSD, we plotted histograms of torsional dissimilarity indices, computed by comparing observed conformations with those of an accepted standard, usually a minimum-energy form. This worked well for simple systems with few conformational variants, but for more complex systems with many conformational possibilities, the projection of multidimensional parameter spaces onto one-dimensional representations led to information loss, due to the overlapping of peaks which arise from different conformers. The small number of structural archetypes for a given metal coordination number suggests that simple histograms of symmetry-minimized $R_c(x)$ values should not suffer from serious information loss and, hence, they should be more valuable in preliminary analyses of coordination sphere geometries than they are in the more complex conformational analyses. However, we note that some degree of information loss is apparent in Figs. 3(*a*) and (*b*), since the close coherence of the T cluster is not well represented in both histograms, nor does the representation of the SQP cluster in one-dimensional space give any indication of the geometrical deformations revealed in the PCA plot of Fig. 2.

Given an awareness of two or more major archetypes for a particular species, either from prior knowledge or from analysis of histograms of the type exemplified in Figs. 3(*a*) and (*b*), it is likely that a two-dimensional scatterplot of the pair of $R_c(x)$ values will provide more information than the one-dimensional histograms alone. Fig. 4 shows the scatterplot of R_c(T) *versus* R_c(SQP), this time for a test dataset of 1306 TrL_4 species, and is clearly intractable to PCA within the *Vista* program. The plot shows a coherent cluster of observations close to R_c(T) of zero, with a clear interconversion pathway leading to a more distorted cluster close to R_c(SQP) of zero. Here, the specific SQP distortion, revealed as a line of density perpendicular to the interconversion pathways in the PCA scatterplot of Fig. 2, can be readily identified in the SQP cluster of Fig. 4. Thus, the two areas of information loss in the one-dimensional histograms, identified above, are both clearly visualized in the two-dimensional $R_c(x)$ scatterplot of Fig. 4.

The $R_c(x)$ analysis essentially assigns each metal coordination sphere to a single asymmetric unit of the six-dimensional valence-angle space, and might be regarded as less complete than the PCA procedure, which treats the complete space. Nevertheless, comparison with the PCA maps shows that the $R_c(x)$ analysis is just as effective in visualizing the geometrical diversity of ML_4 systems, and the $R_c(x)$ values can readily be computed for datasets that cannot be treated by the PCA routines of programs such as *Vista*.

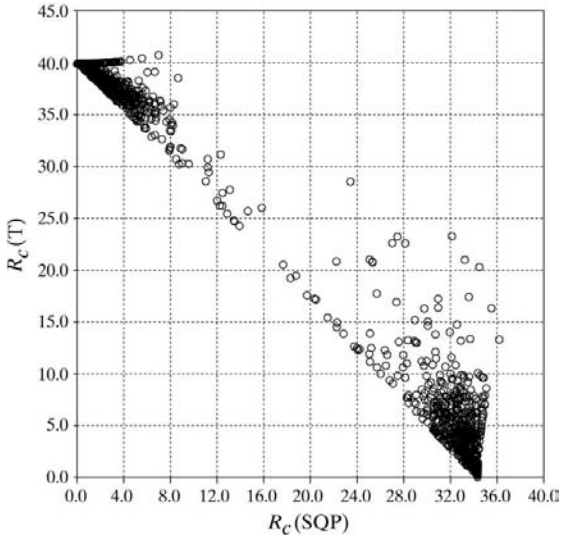


Figure 4
Scatterplot of R_c(T) *versus* R_c(SQP) for the test set of 1306 TrL_4 complexes.

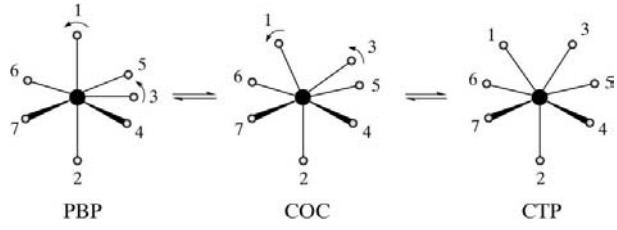


Figure 5
Archetypal polyhedra PBP, COC and CTP for ML_7 coordination. The standard angles for the CTP form used in this analysis are indicated as (°): L_1-M-L_2 141.8, L_1-M-L_3 76.4, L_1-M-L_4 127.3, L_1-M-L_5 127.3, L_1-M-L_6 76.4, L_1-M-L_7 76.4, L_2-M-L_4 76.4, L_2-M-L_5 76.4, L_4-M-L_5 87.8, L_4-M-L_6 152.7, L_4-M-L_7 85.9.

Table 2
Chemical constitution of the test set of 372 ML_7 complexes.

N_e is the No. of structural entries located in the CSD.

M	Sc	Ti	V	Cr	Mn	Fe	Co	Ni	Cu	Zn
N_e	7	27	19	5	27	33	26	4	15	22
M	Y	Zr	Nb	Mo	Tc	Ru	Rh	Pd	Ag	Cd
N_e	13	26	25	95	9	2	0	0	0	33
M	La	Hf	Ta	W	Re	Os	Ir	Pt	Au	Hg
N_e	3	1	30	64	8	3	0	0	0	8

4. Application to ML_7 systems

4.1. Archetypal polyhedra

The three common archetypal polyhedra for seven-coordination (Fig. 5) are the pentagonal bipyramid (PBP, D_{5h}), capped octahedron (COC, C_{3v}), and the capped trigonal prism (CTP, C_{2v}). The COC form can be considered as an intermediate on the PBP/CTP interconversion pathway (Kepert, 1979). In this interconversion, one of the five equatorial ligands in PBP (*e.g.* atom 3 in Fig. 5*a*) moves out of the plane towards the apical ligand 1, which moves away from its apical location. Small relocations of the equatorial ligands 4–7 then generate the COC geometry of Fig. 5(*b*). Further small movements of atoms 3 and 1 then generate the CTP polyhedron of Fig. 5(*c*).

Given the success of the R_c(T) *versus* R_c(SQP) scatterplot for four-coordinate species, it was desirable to generate $R_c(x)$ values for seven-coordinate species referred not only to the PBP archetype, but also to at least one of the other forms. The CTP form was chosen, due to its known interconversional relationship to PBP, and the COC form was an intermediate in this interconversion pathway. However, standard angles for CTP are not fixed by symmetry and were chosen by using the experimental angles from several structures with all unidentate ligands stored in the CSD. The structures chosen had CTP angles that were close to C_{2v} symmetry and had no large chelate effects. Angles within individual structures were first averaged across C_{2v} symmetry and the symmetry-independent angles were then averaged over the chosen structures to generate the θ_{sk} values given in Fig. 5. Further comments on the choice of reference angles are given in §5.

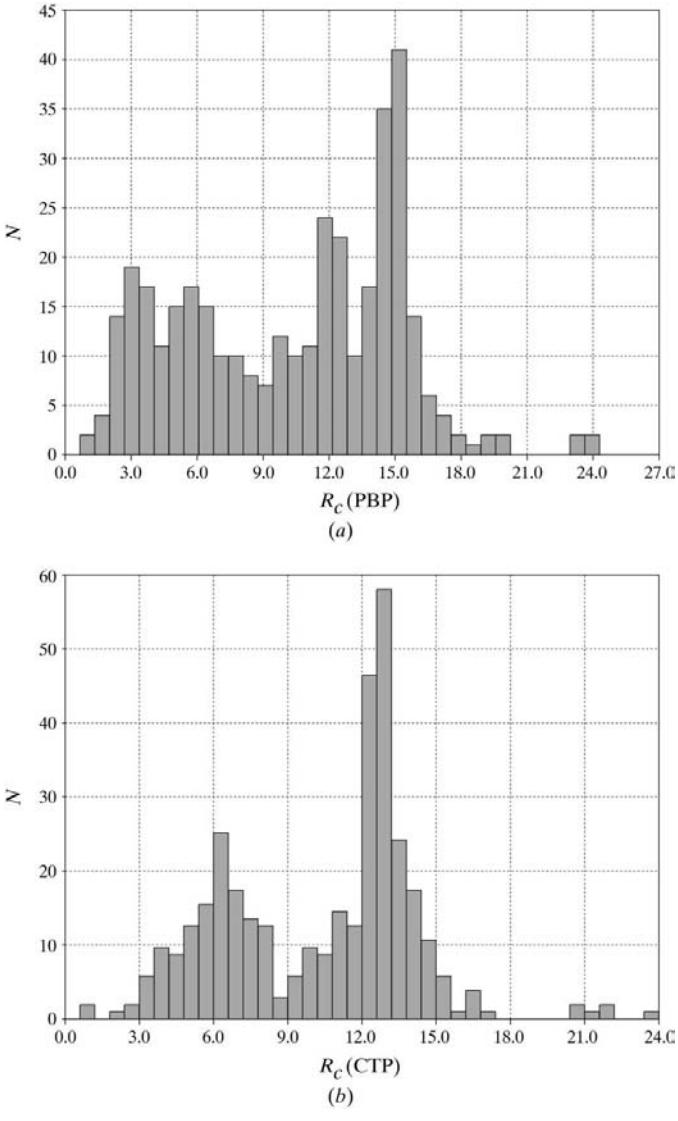


Figure 6
Histograms of (*a*) R_c(PBP) and (*b*) R_c(CTP) for the test set of 372 ML_7 coordination spheres.

4.2. $R_c(x)$ analysis of seven-coordination

4.2.1. Dataset. A test set of 372 ML_7 systems was retrieved from the CSD using the search procedures described in §2. The chemical constitution of this test set is summarized in Table 2, which shows that complexes of Mo^{II}, W^{II} and Fe^{III} are the most highly represented. Only 67 fragments were located (from 62 CSD entries) in which all ligands were unidentate. The permutationally random sets of 21 $L-M-L$ valence angles for each unique coordination sphere were output from *Quest*3D, together with the coordinates of the eight atoms of each ML_7 system.

4.2.2. $R_c(x)$ histograms. Values of R_c(PBP) and R_c(CTP) were computed and minimized over the 5040 possible ligand-atom permutations for each ML_7 system, and histograms of these values are shown in Figs. 6(*a*) and (*b*). Neither histogram shows a peak at $R_c = 0$, indicative of the exact archetypal geometry. Instead, Fig. 6(*a*) has maximum populations in the areas of R_c(PBP) from 3 to 7%, and from 11 to 16%, but with significant density between these maxima. The plot of R_c(CTP) is rather similar, although it might be considered to be more obviously bimodal. However, this may merely be an artifact of the histogram binning process and both histograms indicate two discrete populations: one close to PBP geometry and the other close to CTP geometry, but connected by a significantly populated interconversion region. Histograms for those coordination spheres having only unidentate ligands (not shown) provide clear evidence that PBP geometry is seldom adopted in these cases, with the majority of examples lying closer to CTP geometry or on the CTP–PBP interconversion pathway.

4.2.3. $R_c(x)$ scatterplot. As in the ML_4 case (Fig. 3), the histograms of Fig. 6 provide a valuable initial visualization of

the ML_7 dataset. However, the true value of the $R_c(x)$ analysis is again revealed by the scatterplot of R_c(PBP) *versus* R_c(CTP), shown in Fig. 7, which is extraordinarily rich in information about the geometrical diversity of ML_7 coordination spheres. The plot shows a well populated but elongated cluster of examples (1) having R_c(PBP) in the range 1–6%, and R_c(CTP) in the range 12–14%, representing geometries that approach pure PBP. In the bottom right of the plot, there is a loose cluster of examples (2) in which these $R_c(x)$ values are almost exactly reversed, *i.e.* R_c(CTP) is in the range 1–6%, and R_c(PBP) in the range 14–16%, representing small variations around CTP geometry. This cluster might be expected to be rather loose, since the standard CTP angles (Fig. 5) are not fixed by symmetry. Clusters (1) and (2) are connected by a well populated linear interconversion pathway (A) which passes through cluster (3). Reference to the original data indicates that cluster (3) represents COC geometries.

Clearly, however, geometrical variability in ML_7 systems is not restricted to the PBP–COC–CTP interconversion mechanism illustrated in Fig. 5, and visualized (A) in Fig. 7. This scatterplot also shows a region (4) which has both R_c(PBP) and R_c(CTP) greater than *ca* 14%, and which is associated with two populated pathways: (C) which connects the high-$R_c(x)$ region (4) with the PBP examples, and (B) which connects it to the CTP examples. There is also some indication of a tail of observations (5) for which R_c(CTP) > 15% and R_c(PBP) > 18%.

4.2.4. Colour coding by chemical environment. To further analyse the geometrical variation in seven-coordination, two further chemical dimensions have been added to the $R_c(x)$ plot of Fig. 7 by (*a*) colour coding ML_7 fragments according to the number of unidentate ligands present in each coordination sphere, and (*b*) using the symbols Δ, □ and ○ to indicate the presence of a three-, four- or ≥five-membered chelate ring as the smallest metallacycle in each sphere.

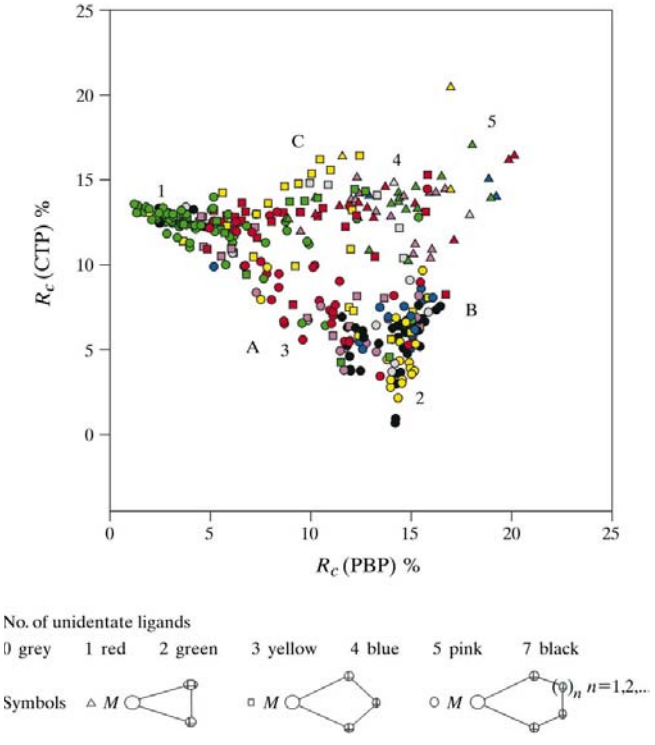


Figure 7
Scatterplot of R_c(PBP) *versus* R_c(CTP) for the test set of 372 ML_7 coordination spheres.

Only six of the ML_7 fragments which have seven unidentate ligands adopt PBP geometries. All have a d^0 configuration and all have at least two halogen ligands, a pair of which occupy axial sites in PBP. This is consistent with molecular orbital calculations (Hoffman *et al.*, 1977), which show that the more electronegative ligands (better σ-acceptors) prefer the apical positions in the PBP geometry. This geometry also tends to minimize ligand···ligand repulsions, which are considered to be a major factor in the choice of geometries in ML_7 systems (Drew & Wilkins, 1974*a*,*b*). In the CTP geometry, the capping position (2 in Fig. 5*c*) is the least crowded and least affected by ligand···ligand repulsions, analogous to the apical positions in PBP. Single halide or other bulky ligands tend to adopt this position in geometries which are very close to CTP and which account for a further 19 fragments. The remaining all-unidentate complexes occur in the CTP–COC area of Fig. 7.

For fragments which do not have three- or four-membered metallacycles (denoted by circles in Fig. 7), those that have two unidentate ligands (green circles) mainly populate the PBP area. These arise principally from Fe complexes that have a pentadentate ligand, occupying the equatorial sites, leaving the unidentate ligands to occupy the apical positions. Fragments having a single unidentate ligand (red circles) are commonly found along the PBP (unidentate ligand in an apical position) ⇔ CTP (unidentate ligand in the capping position 2 in Fig. 5*c*) interconversion pathway, while those with three or four unidentate ligands (yellow or blue circles) populate the COC/CTP area, and have three of these ligands in the apical/capping positions (1, 2 and 3 in Fig. 5*c*). The PBP ⇔ CTP interconversion is depicted in structural terms in Fig. 8.

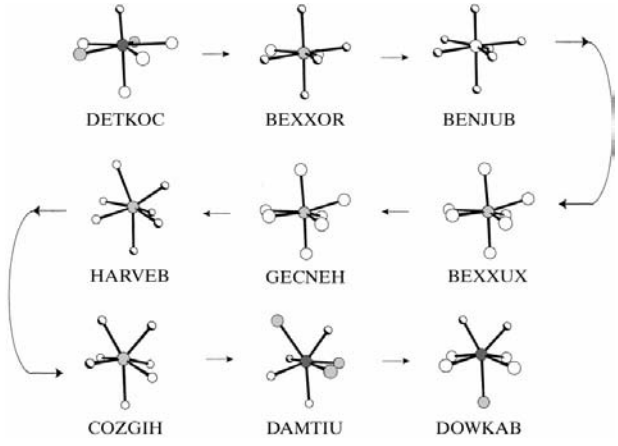


Figure 8
The PBP ⇒ CTP interconversion pathway mapped in structural terms *via* the CSD structures: DETKOC (Blower *et al.*, 1985), BEXXOR (Dewan *et al,* 1982), DENJUB (Beck & Strahle, 1985), BEXXUX (Dewan *et al.*, 1982), GECNEH (Cotton & Matusz, 1987), HARVEB (MacNeil *et al.*, 1993), COZGIH (Hursthouse *et al.*, 1985), DAMTIU (Beauchamp *et al.*, 1985), DOWKAB (Fong *et al.*, 1986).

Examination of the high-$R_c(x)$ area (numbered 4 in Fig. 7) shows that it, and the structural pathways B and C that connect it to the PBP and CTP forms, are almost exclusively populated by structures containing three- and four-membered metallacycles, denoted as Δ and □, respectively. These occupy equatorial sites in both PBP and CTP geometries and the decreased $L-M-L$ bite angles lead to the significantly high $R_c(x)$ values with respect to both archetypes. It is no surprise that the well defined pathway, *B*, that connects PBP forms to the high $R_c(x)$ region (4) is largely populated by four-membered metallacycles close to PBP, while region (4) itself is principally populated by fragments containing three-membered metallacycles, which contain $L-M-L$ angles of $< 60^\circ$.

4.3. Principal component analysis of seven-coordination by selective symmetry expansion

As noted earlier, the $R_c(x)$ minimization procedure locates the closest mapping of the atoms of an observed coordination sphere to the atoms of an archetypal form (x). This means that it establishes a unique enumeration for the atoms of each experimental observation and places it into a single asymmetric unit of valence-angle space. Thus, in order to avoid the full 5040-fold data expansion required to fill that space, we could carry out PCA using just the asymmetric unit identified by the $R_c(x)$ analysis. However, the PC axes chosen are then unrelated to the underlying symmetry of the valence-angle space and as a result the plots are difficult to interpret. Instead, we have chosen to expand the data set in a controlled manner to reflect partial topological symmetries in the resultant maps.

The five equatorial ligands have exact D_{5h} symmetry in the PBP archetype and we have applied a (10-fold) D_{5h} expansion to generate an angular dataset for PCA from the single asymmetric unit identified by $R_c(x)$ analysis. Two datasets have been analysed: (*a*) the 67 fragments that have all unidentate ligands and (*b*) the full ML_7 subset of 372 fragments. In both cases, *ca* 99% (*a*) and 94% (*b*) of the total variance was accounted for by the top ten PCs, with >80% (*a*) and >72% (*b*) of the variance being accounted for by two degenerate pairs, (PC1, PC2) and (PC3, PC4). The scatterplots of PC1 *versus* PC2 for datasets (*a*) and (*b*) are shown in Fig. 9.

Fig. 9(*a*) (all unidentate ligands) shows the small central cluster of PBP examples, corresponding to cluster 1 of Fig. 7. The five symmetry-equivalent and elongated clusters correspond to the COC/CTP region (cluster 2 in Fig. 7), a visual re-expression of the dominance of these coordination geometries for all-unidentate ML_7 coordinations already identified from the $R_c(x)$ analysis. These symmetry-equivalent clusters arise from each of the equatorial atoms of the PBP archetype being designated in turn (by the D_{5h} expansion) as the mobile ligand 3 in Fig. 5. Fig. 9(*b*), mapping the complete ML_7 test sample, now shows the PBP ⇔ CTP interconversion pathway (A in Fig. 7) very clearly, while the COC/CTP area is now much more diffuse by comparison with Fig. 9(*a*). However, the highly distorted cluster 4 of Fig. 7, and the interconversion paths B and C that connect it to the PBP and COC/CTP clusters are more obvious in the $R_c(x)$ scatterplot than they are in the PC1, PC2 scatterplot or in any of the other scatterplots that can be obtained using the five most important PCs from Table 3.

5. Conclusions

This paper has introduced the concept of a Euclidean dissimilarity metric $R_c(x)$ for quantifying the degree of angular distortion of an observed metal coordination sphere from an accepted standard archetype (x). The computational methodology can be applied to any coordination number and has been applied here to example datasets of ML_4 and ML_7 species.

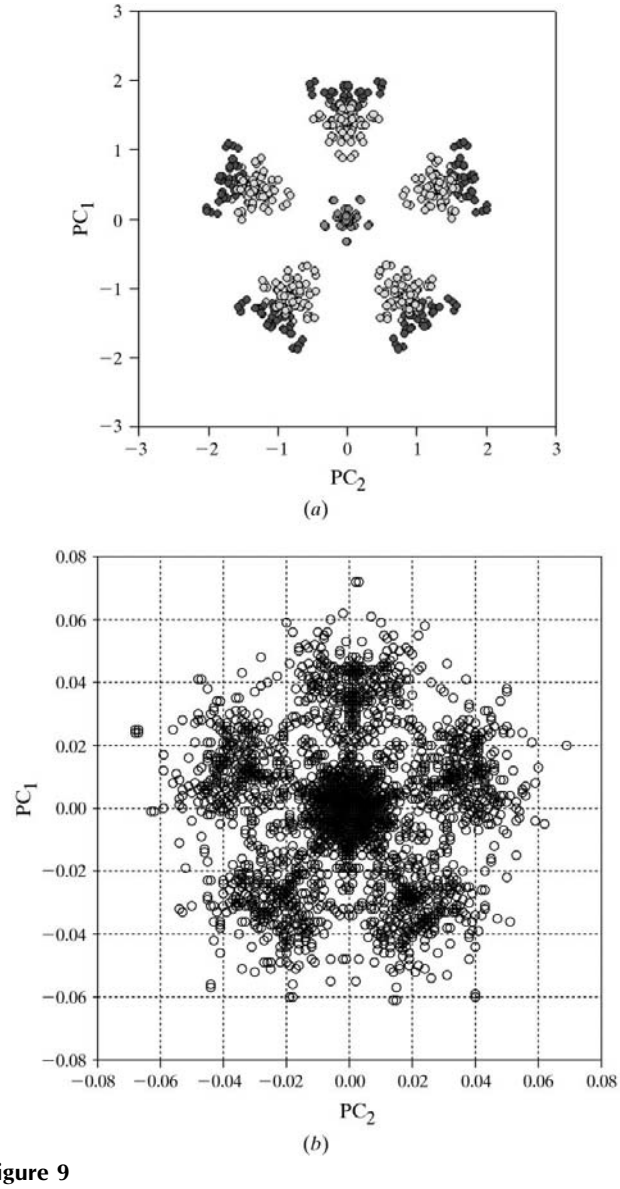


Figure 9
Results of the reduced-expansion PCA for seven-coordination: (*a*) PC1 *versus* PC2 for all-unidentate ligands, (*b*) PC1 *versus* PC2 for all ML_7 examples.

Table 3
PCA results for D_{5h} expanded ML_7 systems: (*a*) all unidentate, (*b*) complete dataset.

Cumul. var. is the cumulative variance.

PC's	PC_1	PC_2	PC_3	PC_4	PC_5	PC_6	PC_7	PC_8	PC_9	PC_{10}
(*a*)										
Variance (%)	22.7	22.7	18.3	18.3	7.5	3.6	3.6	0.7	0.7	0.5
Cumul. var. (%)	22.7	45.4	63.7	82.0	89.5	93.1	96.7	97.4	98.1	98.6
(*b*)										
Variance (%)	19.9	19.9	16.4	16.4	6.7	3.3	3.3	3.2	3.2	1.6
Cumul. var. (%)	19.9	39.8	56.1	72.5	79.2	82.6	85.9	89.1	92.3	93.9

We would stress that the object of the technique is to compare sets of experimentally determined valence angles against one or more sets of reference angles which describe the shape of well known, and preferably easily visualisable, coordination geometries for a specific coordination number. It is most appropriate to choose those reference angles that describe symmetric archetypal forms, even though some of these may not actually be observed in certain classes of experimental structures. Apart from a knowledge of the archetypal forms appropriate to a particular *n*-coordination, we need know nothing about the geometries actually exhibited by real compounds, nor the reasons for their geometrical deviations from the chosen archetypal geometries or from one another. Rather, the technique is data-driven. It is designed to provide low-dimensional mappings relative to well understood reference point(s), which reveal the most important features of a multi-dimensional parameter space, *e.g.* clusters of closely similar geometries, and/or interconversion pathways that traverse that space. The method says nothing about how these mappings should be interpreted, nor about the chemical and structural reasons for the observed distributions of structures. As in all data-mining experiments, that is the task of the experimenter and what we provide here is simply a tool that aids the interpretative process. Thus, it could be argued on chemical grounds that we should have omitted compounds containing three-membered metallacycles from the analysis of seven-coordinate species, since it can be argued that they are a very special case. In fact, we did not know that they were present in the dataset, which included all ML_7 species, but our data-driven procedure has revealed a cluster of structures that do deviate quite markedly from both the PBP and CTP archetypes and were later identified as three-membered metallacycles. Whether these structures remain in a full analysis or not is then entirely at the discretion of the experimenter.

We believe that the analyses described in this paper have demonstrated the richness of information that is contained in simple histograms of $R_c(x)$ and, particularly, in $R_c(x)$ scatterplots in cases where more than one archetypal form (x) exists for a given coordination number. The trials have also shown that the information content of these visualizations can be significantly enhanced by the use of symbols and colour to represent relevant chemical features of individual structures. While analyses of the six-dimensional valence angle spaces generated by ML_4 systems can be, and have been, made using standard multivariate techniques, such analyses become increasingly difficult, if not impossible, to perform as the dimensionality of the problem increases. The $R_c(x)$ analysis has, however, provided one- and two-dimensional visualizations of the 21-dimensional valence-angle space for ML_7 systems that are simple to generate and interpret. The $R_c(x)$ computation also provides the basis for a simplified symmetry expansion of an *n*-dimensional valence angle dataset prior to application of *e.g.* principal component analysis. However, application of this technique to ML_7 systems indicates that the resultant PCA mappings are less amenable to interpretation than the corresponding $R_c(x)$ scatterplot. While further tests are currently in progress for a variety of coordination numbers, initial indications are that $R_c(x)$ values are valuable metrics that are candidates for inclusion in entries in the Cambridge Structural Database. Here, they could form the basis for simple and rapid searches for coordination sphere geometries that are close to, or even far from, an appropriate specific archetypal polyhedron.

The EPSRC(UK) is thanked for the award of a Senior Research Fellowship to JAKH during the tenure of which parts of this work were carried out.

References

Allen, F. H., Bath, P. & Willett, P. (1994). *J. Chem. Inf. Comput. Sci.* **35**, 261–271.
Allen, F. H., Davies, J. E., Galloy, J. J., Johnson, O., Kennard, O., Macrae, C. F., Mitchell, E. M., Mitchell, G. F., Smith, J. M. & Watson, D. G. (1991). *J. Chem. Inf. Comput. Sci.* **31**, 187–204.
Allen, F. H., Doyle, M. J. & Taylor, R. (1991*a*). *Acta Cryst.* B**47**, 29–40.
Allen, F. H., Doyle, M. J. & Taylor, R. (1991*b*). *Acta Cryst.* B**47**, 41–49.
Allen, F. H., Doyle, M. J. & Taylor, R. (1991*c*). *Acta Cryst.* B**47**, 50–61.
Allen, F. H. & Kennard, O. (1993). *Chem. Des. Autom. News*, **8**, 1, 31–37.
Auf der Heyde, T. P. E. & Bürgi, H.-B. (1989). *Inorg. Chem.* **28**, 3960–3969.
Balic Zunic, T. & Makovicky, E. (1996). *Acta Cryst.* B**52**, 78–81.
Beauchamp, A. L., Belanger-Gariepy, F. & Arabi, S. (1985). *Inorg. Chem.* **24**, 1860–1863.
Beck, J. & Strahle, J. (1985). *Z. Naturforsch. Teil B*, **40**, 891–894.
Blower, P. J., Dilworth, J. R., Leigh, G. J., Neaves, B. D., Normanton, F. B., Hutchinson, J. & Zubieta, J. A. (1985). *J. Chem. Soc. Dalton Trans.* pp. 2647–2653.
Bürgi, H.-B. & Dunitz, J. D. (1994). *Structure Correlation*. Weinheim: VCH.
Cambridge Structural Database (1994). *Quest3D User's Manual.* Cambridge Structural Database, Cambridge, England.
Cambridge Structural Database (1995). *Vista User's Guide*, Version 2.0. Cambridge Structural Database, Cambridge, England.
Chatfield, C. & Collins, A. J. (1980). *Introduction to Multivariate Analysis.* London: Chapman and Hall.
Copley, R. C. B. (1995). Ph.D. Thesis, Oxford University, England.
Cotton, F. A. & Matusz, M. (1987). *Polyhedron*, **6**, 261–267.

Dewan, J. C., Wood, T. E., Walton, R. A. & Lippard, S. J. (1982). *Inorg. Chem.* **21**, 1854–1859.
Drew, M. G. B. & Wilkins, J. D. (1974*a*). *J. Chem. Soc. Dalton Trans.* pp. 198–203.
Drew, M. G. B. & Wilkins, J. D. (1974*b*). *J. Chem. Soc. Dalton Trans.* pp. 1654–1661.
Drew, M. G. B. (1977). *Prog. Inorg. Chem.* **23**, 1–207.
Everitt, B. (1980). *Cluster Analysis*, 2nd ed. London: Halstead Heinemann.
Fong, L. K., Fox, J. R., Foxman, B. M. & Cooper, N. J. (1986). *Inorg. Chem.* **25**, 1880–1886.
Hoffman, R., Beier, B. F., Muetterties, E. L. & Rossi, A. R. (1977). *Inorg. Chem.* **16**, 511–522.
Howard, J. A. K., Copley, R. C. B., Yao, J.-W. & Allen, F. H. (1998). *Chem. Commun.* pp. 2175–2176.
Hursthouse, M. B., Thornton-Pett, M. A., Connor, J. A. & Overton, C. (1985). *Acta Cryst.* C**41**, 184–186.
Johnson, M. A. & Maggiora, G. M. (1990). *Concepts and Applications of Molecular Similarity*. New York: John Wiley.
Kepert, D. L. (1979). *Prog. Inorg. Chem.* **25**, 41–144.
Kepert, D. L. (1987). *Comprehensive Coordination Chemistry*, edited by G. Wilkinson, R. D. Gillard & J. McCleverty, Vol. 1, pp. 31–107. Oxford: Pergamon Press.
Klebe, G. & Weber, F. (1994). *Acta Cryst.* B**50**, 50–59.
Lueken, H., Elsenhans, U. & Stamm, U. (1987). *Acta Cryst.* A**43**, 187–194.
MacNeil, J. H., Roszak, A. W., Baird, M. C., Preston, K. F. & Rheingold, A. L. (1993). *Organometallics*, **12**, 4402–4412.
Makovicky, E. & Balic Zunic, T. (1998). *Acta Cryst.* B**54**, 766–773.
Pinsky, M. & Avnir, D. (1998). *Inorg. Chem.* **31**, 5575–5882.
Raithby, P. R., Shields, G. P., Allen, F. H. & Motherwell, W. D. S. (2000). *Acta Cryst.* B**56**, 444–454.
Taylor, R. & Allen, F. H. (1994). *Structure Correlation*, edited by H.-B. Bürgi and J. D. Dunitz, Vol. 1, pp. 111–161. Weinheim: VCH.