Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Geometry of metal–ligand interactions in proteins

Marjorie M. Harding

Institute of Cell and Molecular Biology, University of Edinburgh, Michael Swann Building, Edinburgh EH9 3JR, Scotland

Correspondence e-mail: marjorie.harding@ed.ac.uk

Received 17 September 2000
Accepted 30 November 2000

The geometry of metal–ligand interactions in proteins is examined and compared with information for small-molecule complexes from the Cambridge Structural Database (CSD). The paper deals with the metals Ca, Mg, Mn, Fe, Cu, Zn and with metal–donor atom distances, coordination numbers and extent of distortion from ideal geometry (octahedral, tetrahedral *etc.*). It assesses the agreement between geometry found in all metalloprotein structures in the Protein Data Bank (PDB) determined at resolution ≤ 1.6 Å with that predicted from the CSD for ligands which are analogues of amino-acid side chains in proteins [Harding (1999), *Acta Cryst.* D**55**, 1432–1443; Harding (2000), *Acta Cryst.* D**56**, 857–867]. The agreement is reasonably good for these structures but poorer for many determined at lower resolution (examined to 2.8 Å resolution). For metal–donor distances, the predictions from the CSD, with minor adjustments, provide good targets either for validation or for restraints in refinement of structures where only poorer resolution data is available. These target distances are tabulated and the use of restraints is recommended. Validation of angles or the use in refinement of restraints on angles at the metal atom is more difficult because of the inherent flexibility of these angles. A much simplified set of parameters for angle restraints with quite large standard deviations is provided. (Despite the flexibility of the angles, acceptable and preferred coordination numbers and shapes are well established and a summary table is provided.) An unusual and perhaps biochemically important feature of Zn coordination with carboxylate seen in the CSD examples is also clearly present in metalloprotein structures. With metals like Ca, carboxylate coordination is monodentate or bidentate (two M—O bonds of nearly equal length). In Zn carboxylates a continuous range between monodentate and bidentate coordination is found, with one Zn—O bond of normal length and another of any length between this and a van der Waals contact.

1. Introduction

Two previous papers (Harding, 1999, 2000) have gathered information on metal–ligand geometry relevant to proteins from accurately determined structures of smaller molecules held in the Cambridge Structural Database (CSD; Allen & Kennard, 1993*a*,*b*). The structures were complexes of Ca, Mg, Mn, Fe, Cu and Zn with ligands analogous to those of amino-acid side chains and others found in proteins. The objectives were to provide information for protein crystallography which might be useful (i) in the interpretation and fitting of models to electron-density maps, (ii) for target distances in restrained refinement, (iii) in the validation of protein structural data and possibly (iv) to contribute to the basic understanding of structure in relation to biological function. This paper aims to assess how consistent this geometry derived from the CSD is with the geometry in metalloprotein structures in the Protein Data Bank (PDB; Bernstein *et al.*, 1977; Berman *et al.*, 2000) and then to consider to what extent it could be useful in validation and refinement. First the distances between metal atoms and ligand donor atoms in the metal coordination group are considered, then the size of the coordination groups and their distortions from ideal geometry. The six metals selected are by far the commonest in structures deposited in the PDB. The number of structures determined by diffraction methods and deposited by July 1, 1999 (found using the 3DBrowser and searching for Ca, Mg *etc.* as associated group) was 981 containing Ca, 473 Mg, 225 Mn, 727 Fe, 172 Cu and 649 Zn. 307 structures contained Na; more than 100 structures were not found for any other metal.

The geometry of various kinds of interactions of metal ions with proteins was examined by Chakrabarti (1990*a*,*b*,*c*) in structures available at that time; there were, of course, far fewer than are available now. Further information on metal–protein interactions and geometry can also be found at http://metallo.scripps.edu/.

For any metal–donor atom combination, the bond distance may be affected by the coordination number of the metal, its oxidation state (and spin state) and the chemical nature of the ligand (particularly its charge) and of the other ligands around the metal atom. Metal–ligand distances are 'softer', *i.e.* more variable or flexible, than distances within simple organic molecules and bond-length predictions cannot be as precise as those of Engh & Huber (1991) for the components of a polypeptide chain. Described below is the extraction of metal–donor atom distances for the six metals and their comparison with the distances previously found from the CSD (Harding, 1999). This shows, not surprisingly, that the CSD values are consistent with those in metalloproteins and that the agreement is good for structures determined at high resolution and poorer for low-resolution structures. Information on preferred coordination numbers and ideal geometry of metal complexes is to be found in many textbooks, *e.g.* that of Frausto da Silva & Williams (1991), and is exemplified by the results given by Harding (2000). Interbond angles are 'softer' than bond distances and interbond angles at metal atoms are much softer than those at C atoms; within these rather wide limits the geometry found below in the PDB is in most respects consistent with that in the CSD. Targets for validation or for refinement restraint can be set up well for metal–donor distances but only in a simple form, with large standard deviations, for angles at the metal atom.

2. Methods and procedures

A donor atom L is considered to be in the primary coordination sphere of the metal if it is within (target distance + tolerance) of the metal; a tolerance of 0.5 Å has been found convenient so far. A set of target metal–donor atom distances is given in Table 1. Apart from the values for carboxylate (Asp and Glu) these are the distances found in the CSD and tabulated (Harding, 1999), with some simplification – the use of a weighted average where distances were given there for more than one oxidation state or coordination number. For distances to carboxylate O, CSD values were initially used but were not entirely satisfactory; better values were derived from the PDB results (see later) and are given in Table 1. Target distances were needed for all other possible donor atoms and those used are shown in Table 1(*b*). A secondary coordination sphere was also defined, with metal–donor atom distance between (target distance + 0.5 Å) and (target distance + 1.0 Å). If metal–donor atom distances are found within this range it will usually be because of errors in coordinate refinement or interpretation, but may occasionally indicate significant weak bonding (see Zn–carboxylate binding below and Harding, 2000).

Table 1
Metal–donor distances.

(*a*) Metal–donor atom 'target' distances (Å). Apart from the values for carboxylates (Asp and Glu), these are based entirely on distances found in the CSD. The values given for Asp and Glu are for monodentate carboxylate; in bidentate carboxylates the values are increased by 0.11 Å; for Ca, Mg, Mn and Zn the values were derived from the analysis of the PDB results (with resolution ≤1.6 Å), where many more observations were available than in the CSD. Fe and Cu carboxylate values are based on the CSD. All the values for Cu^{II} are based on the short bonds found in four-coordinate Cu^{II} and in the basal plane of five- and six-coordinate Cu^{II}.

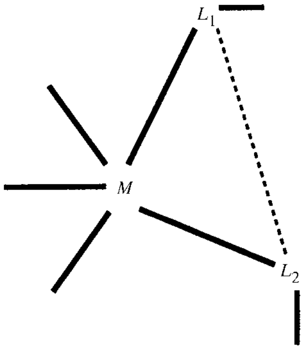
Values in parentheses are estimates only; there is insufficient data in the CSD.

	O water	O Asp or Glu, monodentate	O Ser or Thr	O Tyr	N His	S Cys
Ca	2.39	2.36	2.43	(2.20)	(2.38)	(2.56)
Mg	2.07	2.26	2.10	(1.87)	(2.05)	(2.23)
Mn	2.21	2.21	2.25	1.88	2.19	2.35
Fe	2.09	2.01	2.13	1.93	2.08	2.27
Cu^{II}	1.97	1.96	2.00	1.90	1.99	(2.17)
Zn	2.09	2.04	2.14	1.95	2.00	2.29

(*b*) Target distances used for other metal–donor atom combinations.

M–main-chain carbonyl O	As $M-$O Asp (monodentate) above
M–any other O donor	As $M-$O Ser above
M–any other N donor	As $M-$N His above
M–any other donor	As $M-$S Cys above

The full PDB files used for coordinate extraction were those released by RCSB (1999). A local program (*MP*) has been written which reads a PDB file, extracts information from the header and extracts the coordinates and occupancies of each metal atom in the structure and those of every atom other than C, P or H within 3.6 Å of the metal atom. Target distances for each metal–donor combination are looked up from a table and compared with the observed distances in order to list all contacts within the primary and then the secondary coordination spheres. This allows evaluation of coordination numbers (primary and secondary). Interbond angles, L_1-M-L_2, are then calculated for the primary coordination sphere and, in the case of four-, five- or six-coordinate metal centres, the r.m.s. deviation of these angles is calculated from the ideal ones for the possible geometries (Harding, 2000). For each structure, two files may be output: the first contains selected information from the header and the contact and geometry details; the second contains coordinates of the metal, the donor atoms *and* all the atoms of the amino-acid residue or ligand to which each donor belongs, in a format suitable for input to the graphical display program *PLUTO* (a part of the the CSD system).



The PDB file normally contains one asymmetric unit of the crystal structure. The program *MP* does not (yet) generate any symmetry-related atoms; thus, if the metal atom lies on a crystallographic rotation axis the coordination number found by *MP* is incorrect and low. A third form of output from the program *MP*, developed after the main survey was performed, is a .bccab file acceptable as input to the program *PREQUEST* in the CSD system for the building of a 'private' database which can be queried by other CSD programs. This option was used and the metal coordination number established through *QUEST* for all examples where it appeared to be lower than expected. However, contacts to a metal atom in a general position from donor atoms in neighbouring asymmetric units related by crystallographic symmetry are still not identified.

Lists of PDB codes were obtained using the PDB 3DBrowser at the mirror site http://www.ebi.ac.uk/pdb/index.shtml and searching for structures determined by diffraction methods within the resolution range required and containing Ca, Mg, Mn, Fe, Cu or Zn as 'associated group'. (These lists include some nucleic acid structures and exclude some specially named metal-containing 'het-groups'.) The program *MP* was then run for series of, for example, 100 or more Ca-containing proteins, giving a log file containing the identity, contact and coordination group information for each (elapsed time on an SG workstation less than 1 min). This log file was then quickly converted to two files suitable for input to the spreadsheet program *VISTA* (a part of the CSD search and analysis system); the first file contained one entry per metal–donor atom contact and the second contained one entry per metal centre. In the analysis of interatomic distances, contacts involving disordered atoms were excluded; in the analysis of coordination numbers, metal coordination groups including disordered atoms were excluded. Using *VISTA*, distributions were examined, mean and standard deviations obtained and outliers and anomalies identified. Where the anomalies involved atoms of low occupancy they were excluded. Outliers and anomalies could be further investigated using the contact information file (which also included *B* values of the atoms) or the graphics file output by *MP* for individual structures.

For the main comparisons of metal coordination geometry in the PDB, all structures determined at resolution $\leq$1.6 Å were searched for metal atoms. The quality of metal coordination geometry reported in structures determined at lower resolution ($\leq$2.8 Å) was then explored, but using only the subset 'representative macromolecules' – this gave adequate numbers for useful comparisons but excluded structure determinations of closely related proteins such as mutants. A small number of protein structures were excluded on account of format errors in the PDB files, more than 20 000 atoms in the asymmetric unit (exceeding the program capacity) and similar reasons.

3. Results and discussion

3.1. Contact distances in structures with resolution $\leq$ 1.6 Å

Geometrical information on 390 metal centres in 229 metalloprotein structures was retrieved in this search. Tables 2(*a*) and 2(*b*) give numbers of observations and mean dif (dif = observed distance − target distance) for the common amino-acid side-chain donors and for the main-chain carbonyl group, which it was found desirable to include. Table 2(*c*) gives information on the other contacts found.

The division of the metal–donor atom distances into primary and secondary coordination spheres and the tolerance values chosen to do this appear satisfactory. Table 2(*a*) and the distributions of dif (examples in Fig. 1) show that in most cases the number of contacts in the secondary coordination sphere is very small and that the mean values of dif are small compared with their sample standard deviations (which represent the scatter of the observations). Apart from some carboxylates (see later), contacts assigned to the second coordination sphere are likely to represent errors in interpretation or in parameter refinement. The mean values of dif may not be significantly different from zero; however, non-zero values could arise if the mix of ligand types or coordination numbers in the PDB structures is different from that in the complexes of the same metal with the same donor in the CSD. The standard deviations of dif, mostly 0.10–0.15 Å, and the scatter of values in the PDB results are such that expected variations in metal–donor distances with coordination number cannot be reliably detected.

More critical comparisons with some of the target distances were achieved by selecting only the protein structures determined at a resolution of $\leq$1.2 Å. The total number of observations for most of the ten metal–donor combinations in this set is small, but overall there are 261 observations with a mean dif of 0.01 (5) – details are given in Table S1, deposited as supplementary information.[1]

---

[1] Supplementary material has been deposited in the IUCr electronic archive (Reference: ad0141). Services for accessing these data are described at the back of the journal.

3.2. Effects of resolution, uncertainties

Excluding Cu, for which there are special problems, the average sample standard deviation for the target distances from CSD given in Table 1 is 0.06 Å and the maximum is 0.10 Å. These structure determinations have used diffraction data to a resolution of 0.9 Å or better. The small group of protein structure determinations at resolutions ≤1.2 Å have achieved a precision approaching this. For all metalloprotein structures with resolution ≤1.6 Å the mean values of dif have a sample standard deviation averaging 0.12 Å.

In order to see how errors in contact distances changed with resolution, contact distances were extracted and analysed for structures with resolution ≤2.8 Å. The number of structures in this resolution range is large, so the search was restricted to the PDB category 'representative macromolecules', which

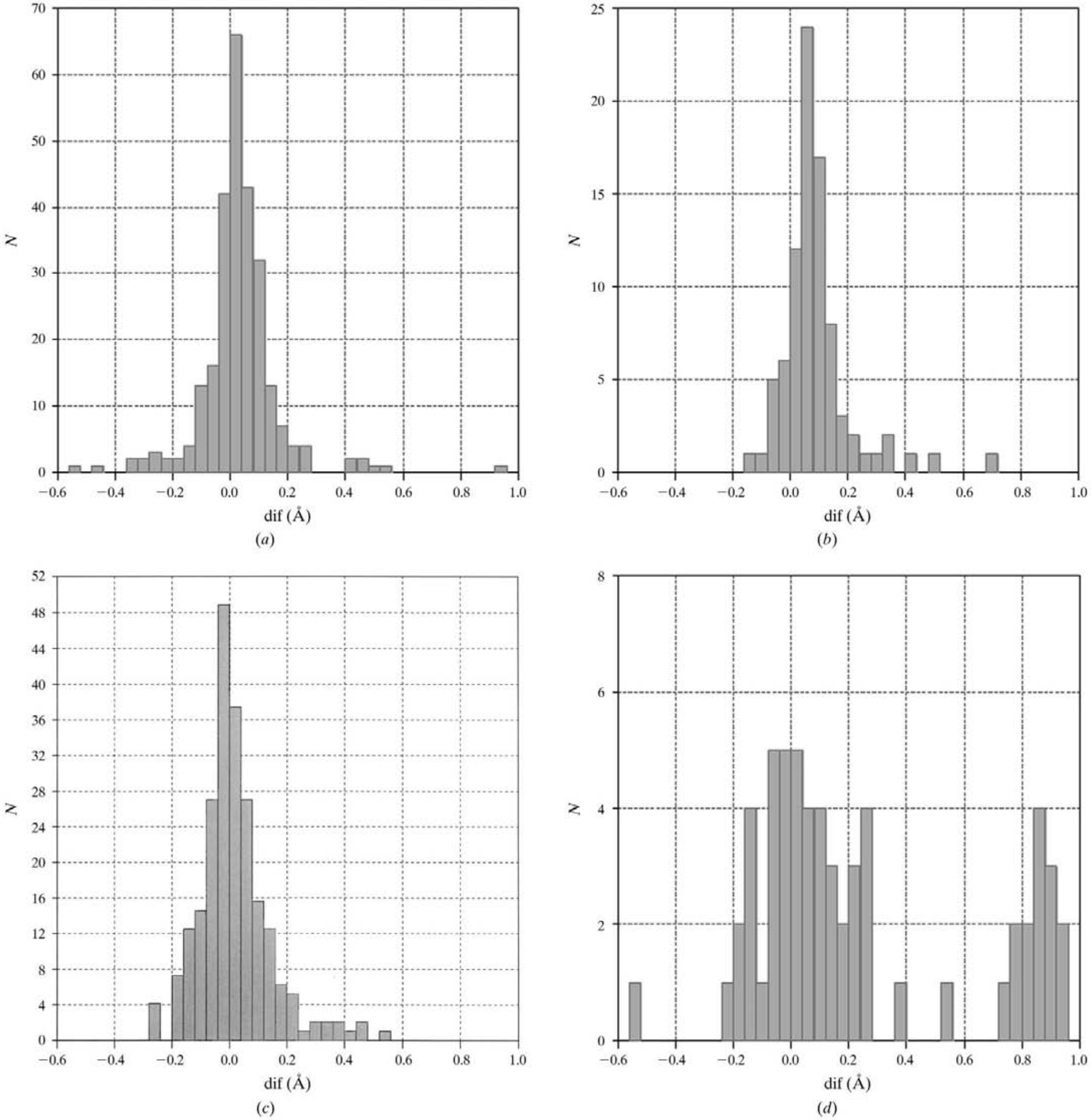


Figure 1
Distribution of dif = observed distance − target distance (in Å) for structures determined at resolutions ≤1.6 Å. (*a*) Ca—O_{water}, (*b*) Zn—$N_{histidine}$, (*c*) Ca—$O_{carboxylate}$, (*d*) Zn—$O_{carboxylate}$. In (*a*)–(*c*) the number of observations with dif outside ±0.5 Å is very small, but in (*d*) it is substantial.

Table 2
Results from structures in PDB with resolution ≤1.6 Å.

(*a*) Numbers of contact distances in the first coordination sphere (*i.e.* within ±0.5 Å of target distance) and, in parentheses, the number of additional contacts in the second coordination sphere out to 1.0 Å. (There are only two contacts to Tyr altogether, from Fe.)

		O Asp or Glu					
	O water	Mono-dentate	Bi-dentate	O Ser or Thr	O carbonyl (main chain)	N His	S Cys
Ca	261 (3)	93 (1)	127	23	143	0	0
Mg	211 (7)	24 (1)	4 (2)	10	6	2	0
Mn	9	17	17 (2)	0	0	5	0
Fe	21 (1)	13	0	0	2 (12)	116	54
Cu	2	0	0	0	0	39	14
Zn	26 (3)	21 (16)	23	2	3 (8)	85 (1)	39

(*b*) Mean value of dif (in Å), where dif is (observed distance − target distance) and the sample standard deviation of dif for contacts within the first coordination sphere.

		O Asp or Glu					
	O water	Mono-dentate	Bi-dentate	O Ser or Thr	O carbonyl (main chain)	N His	S Cys
Ca	0.03 (12)	0.00 (10)	0.02 (13)	0.09 (9)	0.00(11)	—	—
Mg	0.05 (11)	0.00 (11)	0.00 (4)	0.12 (9)	0.23 (12)	0.4	—
Mn	0.13 (16)	0.02 (9)	−0.08 (10)	—	—	0.15 (15)	—
Fe	0.09 (12)	−0.04 (13)	—	—	0.13	0.07 (8)	0.00 (6)
Cu	0.36	—	—	—	_	0.04 (8)	−0.03 (4)
Zn	0.02 (18)	0.05 (14)	0.09 (14)	0.15	0.11	0.08 (10)	0.02 (9)

(*c*) Other M—O and M—N contacts, numbers and mean dif as defined in (*a*) and (*b*). There were also 14 Fe—S contacts with mean dif 0.06 (10) Å.

	Metal–oxygen		Metal–nitrogen	
	Number	Dif	Number	Dif
Ca	33 (1)†	−0.06 (7)	2 (1)	−0.05
Mg	46 (5)	0.14 (16)	2 (3)	0.25
Mn	6	0.08 (7)	—	—
Fe	7 (4)	0.01 (3)	329 (1)‡	−0.06 (5)
Cu	1	0.21	—	—
Zn	5 (5)	−0.05 (6)	6 (1)	0.13 (13)

† Most of these are Asn or Gln. ‡ Most of these are haem N.

gave an adequate sample. Results are shown in Fig. 2 and Table 3. It is not generally clear whether restraints to the geometry around the metal atom have been used in these refinements, but it seems probable that in many cases they have not. Many reported distances differ widely from the chemically predicted ones and in a random way (*i.e.* positively and negatively). The differences increase substantially as $d_{\min}$ increases and they do so roughly as one would expect errors to increase in accordance with Cruickshank's diffraction precision indicator (DPI; Cruickshank, 1999). Using the examples given by Cruickshank and treating the standard deviation of an M—O bond length as up to 1.2 times that of one atom (C, N, O), we could expect a standard deviation of *very roughly* 0.08 Å at 1.2 Å, 0.2 Å at 1.6 Å and 0.5 Å at 2.2 Å.

The refinements of very high resolution data ($d_{\min} \leq 1.2$ Å) have mostly been performed with *SHELXL* (Sheldrick & Schneider, 1997) and with no restraints on metal coordination geometry. At this resolution, the diffraction data can determine the metal–donor atom distances as well as or better than they can be predicted. In the poorer resolution studies, certainly in the 2.4–2.8 Å range and probably from ~1.6 Å upwards, diffraction data is clearly unable to determine metal–donor distances to the precision that they could be predicted and it would be highly appropriate to apply restraints.

3.3. Comments on carboxylate coordination, special behaviours of Zn

Carboxylate groups show more complications than do other ligands. A carboxylate group may coordinate to metal through one O atom or through both O atoms, or the two O atoms may be coordinated to two different metal ions and the carboxylate described as bridging. The M—O bond to a bidentate carboxylate is expected to be slightly longer than that to a monodentate carboxylate. For Zn complexes in the CSD a continuous range of states occurs between the monodentate and bidentate coordination of carboxylate to one metal ion (as well as bidentate bridging two metal ions); monodentate coordination has Zn—O1 = 1.95 (7) Å, Zn—O2 > 3.2 Å, bidentate has Zn—O1 ≃ Zn—O2 ≃ 2.2 Å and in the intermediate states there is an inverse correlation between Zn—O1 and Zn—O2 (Harding, 1999, 2000). For the other metals, the situation is not so clear in the CSD and the number of observations is often rather small. For Mg and Mn there are more observations in the PDB (with resolution ≤1.6 Å) than in the CSD and many more for Ca. For Mn in the PDB, as well as Ca in the CSD and PDB, there are clear examples of bidentate coordination with both M—O contacts in the first coordination sphere; the average M—O$_{\rm bidentate}$ distance is longer than the average M—O$_{\rm monodentate}$. (For Mg, there are two examples of bidentate carboxylate groups, both in the same structure.) There is no strong evidence for a range of intermediate states like those found for Zn, with M—O distances in the second coordination sphere. Zn–carboxylate interactions in the PDB do appear to exhibit a range of intermediate states similar to those found in the CSD. As well as 23 examples of Zn bonds to the O atom of bidentate carboxylate, there are 16 examples where the carboxylate group has one O atom in the first coordination

Table 3
Variation of agreement with resolution, for 322 metalloprotein structures in ‘representative macromolecule’ set in PDB and these include 735 metal centres..

Resolution range (Å)	0–1.6	1.61–2.2	2.21–2.8
Total number of contacts	375	1681	877
Mean ‘dif’ in first coordination sphere (Å)	0.03	0.03	0.04
% of contacts found in second coordination sphere			
Excluding Zn	1	5	16
All metals	4	6	14
Mean sample standard deviation for ‘dif’ (Å)			
First sphere	0.11	0.16	0.19
First and second spheres (Å)	0.18	0.23	0.28

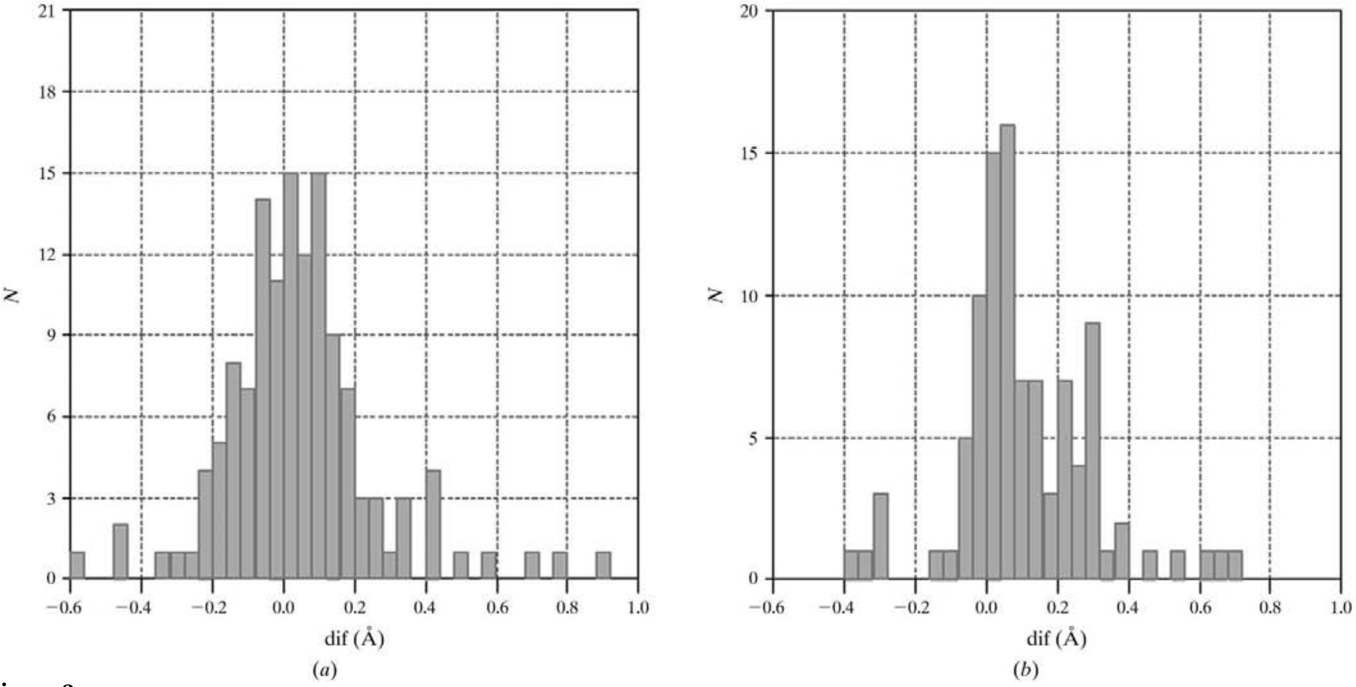
sphere around Zn and one O atom in the second sphere. The distances follow the pattern shown in Harding (2000), although somewhat roughly, since the errors are larger (Fig. 3). Thus, the large number of Zn–carboxylate contacts which appear to be in the second coordination sphere need not be attributed to coordinate errors. This behaviour may well be of significance in the catalytic function of Zn enzymes.

In the setting of target values for carboxylates from the CSD, a constant difference between the monodentate and bidentate carboxylates was assumed and was the same for all metals. The targets initially set for Ca, Mg, Mn and Zn were not entirely satisfactory and were revised (to the values given in Table 1) to give better agreement with the PDB results, in which many more observations are available for these metals. (The Mg—O$_{\text{monodentate}}$ distance was increased by 0.15 Å; all other changes were ≤0.08 Å in magnitude.) Thus, Table 2 shows very good agreement for the carboxylates of Ca, Mg and Mn. For Zn, the larger values of dif, standard deviation and the number of observed contacts in the second coordination sphere are the result of the flexibility of the carboxylate coordination described above.

3.4. Comments on contact distances for other individual metals and donors

3.4.1. Main-chain carbonyl O. Main-chain carbonyl O as a donor is common with Ca, but uncommon with other metals. A search in the CSD for analogous M—O distances using a search fragment M—O═C—N, with C and N constrained to have sp^2 coordination and O to have no other atoms bonded to it, yielded mean distances for Ca—O of 2.35 (6), Mg—O of 2.07 (3) and Zn—O of 2.05 (9) Å from 63, 55 and 115 observations, respectively, consistent with the PDB observations. Many different amino-acid residues participate in such Ca—O bonds; the commonest in the ≤1.6 Å resolution set are Val (22 examples), Gly (16), Asn (16), Thr (15) and Tyr (14). These relative frequencies may be affected by the presence of repeat structure determinations of very similar proteins (mutants, derivatives *etc.*) In the structures of ‘representative macromolecules’ at resolutions ≤2.8 Å, the commonest are Gly (66 examples), Thr (35), Ala (27), Asp (27) and Asn (20), with 14 other amino acids participating in 2–14 examples each.

3.4.2. Coordination by other amino-acid side chains. There are in this set of proteins 16 examples of Asn coordination to Ca and ten of Gln coordination to Ca, with a mean Ca—O distance for Asn and Gln of 2.32 Å, and 11 cases of Met coordination to Fe, with a mean Fe—S distance of 2.34 Å. There are 13 examples of Met coordination to Cu, with Cu—S distances in the rather wide range 2.5–3.1 Å. There are five



(*a*) (*b*)

Figure 2
Distribution of dif = observed distance − target distance (in Å) for structures determined at resolutions 1.61–2.2 Å. (*a*) Ca—O$_{\text{water}}$, (*b*) Zn—N$_{\text{histidine}}$. These should be compared with Figs. 1(*a*) and (*b*); here, as expected, the spread of values of dif is greater.

examples of other amino acids coordinated to these metals, two of which are N of Lys to Zn at 2.13 Å.

3.4.3. Coordination of metal by other ligands. A variety of organic ligands with N or O donor atoms are found and the target distances used for these (Table 2*b*) have proved satisfactory and also for Fe—S bonds to iron–sulfur clusters. Zn—Cl and Cu—Cl bonds appear in small numbers.

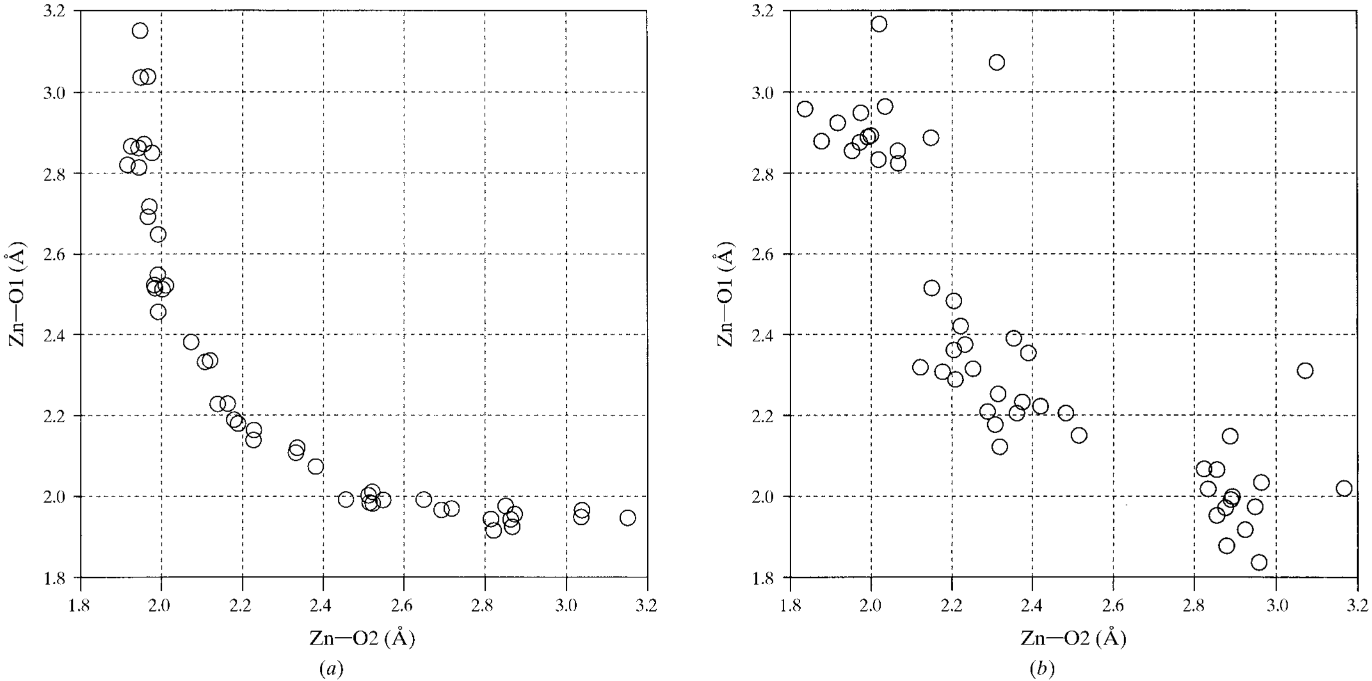
3.4.4. Complexes of Cu. Target distances have not yet been derived from the CSD for Cu^I distances; there are very few. Cu^{II} complexes are difficult on account of the Jahn–Teller effect. Normally, there should be four short well predictable Cu–donor bonds, *e.g.* of 1.97 (3) Å when H_2O is the donor, in a square-planar arrangement. In addition, there may be one or two additional bonds, making the complex a tetragonal pyramid or octahedron; these are more variable in length, *e.g.* 2.30 (9) or 2.42 (16) Å for H_2O as fifth or fifth and sixth ligand – the length depends on the ligand field strength of the other ligands around the Cu^{II} and the large standard deviations reflect the wide range of ligand types present in the different complexes in the CSD.

3.5. Suggested target distances for refinement restraints and for validation

Metal–donor atom distances depend on the ligand and also on the metal-atom oxidation state, coordination number and, to a small extent, on the other ligands around the metal. If values for distances are to be correctly predicted with a standard uncertainty (s.u.) approaching 0.05 Å, all these factors must be matched in the model compounds used for the predictions. In this case, target distances should be taken from Harding (1999) and even there some oxidation states or coordination numbers or ligand types are not adequately represented. [Better still, of course, will be detailed comparisons or reviews for particular types of coordination group, for example Parisini *et al.* (1999) on Fe_4S_4 clusters, Maher *et al.* (1999) on FeS_4 groups as in rubredoxin, Chong *et al.* (1999) on haem groups in deoxy- and carbonmonoxy-haemoglobins.]

If target distances with an s.u. of 0.10 Å are acceptable, then the values in Table 1 are recommended, with the exceptions below. These would be applied as restraints of the type DFIX in *SHELXL* (Sheldrick & Schneider, 1997), which uses both distance and standard deviation. For Zn carboxylates, one carboxylate O must be within the first coordination sphere, with Zn—O = 2.09 (10) Å; for the other carboxylate O, Zn···O may take any value greater than 2.09 Å. [Alternatively, a constraint of the following form could be used when both Zn···O distances are less than 3.0 Å; it has been devised by fitting observed distances in simple $(RCO_2)_2ZnX_2$ complexes. For Zn···O in Å, $(\mathrm{Zn—O_1} - 1.78) \times (\mathrm{Zn—O_2} - 1.78) = 0.168$.] Four-coordinate Cu^{II} complexes could be restrained to square-planar geometry with the distances given in Table 1 with a standard deviation of 0.1 (or even 0.05) Å; until the coordination number and the basal plane can be identified with certainty, the best restraint that can be given for five- or six-coordinate Cu^{II} with O ligands is to 2.1 (3) Å.

For validation, the most useful check is that the reported bond distance should be within 0.5 Å of the target distance (and it would be adequate to use one target distance for all contacts of one metal to oxygen, *e.g.* the distance to Ser O). Fairly high resolution structures (better than 1.6 Å) which fail this test are uncommon, except for those with zinc coordinated to carboxylate, and need to be examined for errors in interpretation or refinement. In one example (PDB code 1lam), the



(*a*) (*b*)

Figure 3
$\mathrm{Zn—O_1}$ and $\mathrm{Zn—O_2}$ distances (in Å) for carboxylate groups coordinated to Zn. (*a*) in simple complexes, $(RCOO)_2ZnX_2$ in the CSD and (*b*) in Zn proteins in the PDB with structures determined at resolution ≤1.6 Å.

Table 4
Summary of coordination number (CN) and shape of coordination group, based on structures in the CSD with ligand donor atoms N, O, S or Cl.

Shapes are tetrahedral (tet), square planar (sqp), trigonal bipyramidal (tbp), tetragonal pyramidal (tetp) and octahedral (oct). δ for a metal complex is the r.m.s. deviation of the angles around the metal atom from the ideal values for the specified geometry (see Harding, 2000, for full details).

	4		5		6		
Coordination number	Non-chelated	Chelated	Non-chelated	Chelated	Non-chelated	Chelated	Other
Mg	Few	Some	Very few	Some	Common	Common	Occasionally CN = 7 with very highly chelating ligands
Mn	Some	Some	Very few	Common	Common	Very common	
Fe	Common	Common	Very few	Common	Some	Very common	
Cu$^{\rm I}$	Some	Common	—	Few	—	—	CN = 3 quite common
Zn	Common	Common	—	Common	Some	Common	—
Preferred geometry (possible geometry)	tet	tet (sqp)	(tetp, tbp)	tetp (tbp)	oct	oct	
Mean δ (°)	4	10	(4)	6	2	8	
Cu$^{\rm II}$	Common	Very common	Few	Very common	Some	Common	—
Preferred geometry (possible geometry)	sqp (tet)	sqp (tet)	(tbp, tetp)	tetp (tbp)	oct	oct	
Mean δ (°)	12	8	3	7	2	8	
Ca	—	—	Very few	Very few	Common	Common	CN = 7 common
Preferred geometry (possible geometry)	—	—	(tetp, tbp)	(tetp, tbp)	oct	oct	CN = 8 common with chelating ligands
Mean δ (°)	—	—	(6)	(16)	5	16	

PDB coordinates indicate a Zn atom coordinated to four main-chain carbonyl O atoms at 2.5–2.9 Å, one water molecule at 2.7 Å and no other donors. This is a decidedly abnormal situation; reference to the original paper (Straeter & Lipscomb, 1995) shows that the authors were fully aware of this and have attributed an unexplained peak as zinc although it might be Na or K. However, there are a small number of other examples in the PDB where fairly high resolution structures with highly improbable contact distances have been deposited and there is no comment in any published paper or in the PDB header – a comment in the PDB header is to be strongly recommended, indicating what steps have been taken to check the interpretation and refinement in this region or any explanation proposed. In lower resolution structures, improbable contact distances are likely to be the result of limitations in the refinement and it is here that the use of restraints would be desirable. The restraints recommended above should bring all distances within an acceptable range. If a separate target distance is used for the different types of M—O distance, then the acceptable range for observed distances should be target ± 0.3 Å.

3.6. Coordination group shape and distortion of angles

Table 4 summarizes the expectations about coordination number and shape based on the analysis of metal complexes with N, O, S or Cl donor ligands in the CSD – and in agreement with a large body of previous chemical experience! Ca and Cu$^{\rm II}$ are treated separately: because of the larger size of Ca^{2+} the coordination numbers in its complexes are larger than those for the other metals; because of its electronic structure, four-coordinate Cu$^{\rm II}$ has different geometry and much more flexible geometry than the other metals.

Any account of the shape of the coordination group must take account of whether the ligands are chelating ligands or not. Multidentate chelating ligands can force geometry that is rare or impossible in non-chelated complexes; even bidentate ligands can force angles to take values well away from their optimum values; for example, four-coordinate zinc is normally tetrahedral, but can be square planar when bound to a porphyrin ligand. The chelating ligands of these small-molecule complexes are molecules in which donor atoms are linked to each other through a small number of bonds, usually 2, 3 or 4. The bidentate carboxylate group is the only chelating ligand occurring in the amino-acid side chains of normal proteins; if bonded to Mg the angle between the two Mg—O bonds is ~60°, if bonded to Ca the angle between Ca—O bonds is ~50°. Either requires substantial distortion away from the ideal octahedral angle of 90° and even for seven-coordinate Ca geometry the angles must be distorted. Many other chelating situations are found in metalloprotein structures when substrate analogues or other ligands are present and for all these it is reasonable to expect ranges of coordination number and angles like those found in the *chelated* small-molecule complexes. A metalloprotein structure in which two or more amino-acid side chains have donor groups coordinated to the metal is formally chelated, but in this case there are many bonds, usually ten or 20 or more in a somewhat flexible chain, between the donor atoms. The distortion of angles around the metal atom to accommodate this chelating ligand may therefore not need to be as great as in simpler chelated small-molecule complexes.

Table 5 shows the coordination numbers found in metalloprotein structures in the PDB in all structures determined with resolution ≤1.6 Å and their average distortions from ideal geometry. With a few exceptions, they conform well to the expectations based on the CSD. As expected, large distortions are found for six-coordinate Ca compared with other ML_6 and even larger distortions are found for four-coordinate Cu. All cases where the coordination number (CN) appeared to be abnormal (Ca, CN < 5 or > 8; other metals, CN < 4 or > 6) were examined. One is a three-coordinate Cu^{I} site (PDB code 1jcv), which is acceptable. The one example of seven-coordinate Fe has both O atoms of an oxygen molecule coordinated to a haem Fe, *i.e.* it is acceptable as a highly chelated situation. There are 11 examples of Cu^{II} proteins, including plastocyanins and azurins (*e.g.* PDB codes 1paz, 1plc, 1pnc, 1pnd; all are related to oxidation/reduction processes) with very similar geometry including three donors (Cys or His) within the first coordination sphere and a long bond to Met in the second coordination sphere, Cu—S = 2.5–3.0 Å; clearly, this is a biologically significant motif. In five other cases (two Ca, three Zn), inclusion of the second coordination sphere donors makes the coordination number acceptable; either there is distortion from normal bond distances or errors in interpretation/refinement. For one very small Zn protein (PDB code 1ppt) the PDB file header explains that the coordination sphere of Zn is completed by other donor groups from protein molecules in neighbouring asymmetric units. For the remainder (three Ca centres with CN < 5; nine Zn centres with CN < 4) the PDB headers give no clues; there is the possibility, not tested, of other donors in neighbouring asymmetric units, but a more likely explanation is simply failure to identify solvent/water molecules present, perhaps disordered ones.

Coordination numbers in the metalloprotein structures determined at lower resolution (1.6–2.8 Å) were also examined. Here there was clearly much more variability of coordination number, possibly because the less reliable metal–donor distances left some donors apparently outside the first coordination sphere and possibly because of donors (*e.g.* water O) not yet identified.

3.7. Suggestions related to interpretation, refinement or validation of coordination group shape and angles

The first and most important check to be made is that the number of atoms in the first coordination sphere (target distance + 0.5 Å) falls within the accepted range for the metal. In map interpretation, consideration of the *preferred* coordination numbers is desirable. Except for Zn and Cu^{II}, it is not normal to find other donor atoms in the second coordination sphere (target distance + 0.5 to target distance + 1.0 Å). In the validation of poorer resolution structures, where metal–donor atom distances have not been restrained, there are likely to be disagreements. Secondly, the coordination shape should conform roughly to that expected, *e.g.* Zn is normally tetrahedral, not square planar.

Table 5
Metalloprotein structures in the PDB with structures determined to resolution ≤1.6 Å.

(*a*) Observed coordination numbers. For structures where the metal atom lies on a crystallographic rotation axis, the coordination number has been adjusted to take account of this. Structures with very low metal-atom occupancy have been excluded, as have DNA structures.

Coordination number	<4	4	5	6	7	8
Ca	3	2	8	23	53	17
Mg	—	—	10	48	—	—
Mn	—	—	—	8	—	—
Fe	—	12	54	41	1	—
Cu	13	4	4	3	—	—
Zn	12	40	12	5	—	—

(*b*) Average deviations, δ (°), from ideal geometry. For coordination number 4 the deviation is from tetrahedral, except for Cu, where it is from square planar. For coordination number 5 the deviation is from tetragonal pyramidal (tetp, the commoner) or trigonal bipyramidal (tbp), whichever is smaller. For coordination number 6 the deviation is from octahedral.

Coordination number	4	5	6
Ca	—	15	14
Mg	—	10	8
Mn	—	—	8
Fe	4	2.5†	6†
Cu	17	18	—
Zn	11	13	—

† Dominated by large number of haem proteins.

Restraints in refinement of the values of angles between metal–donor bonds are more difficult, on account of the flexibility of these angles. As with metal–donor atom distances, the application of restraints is only worthwhile when the precision with which the angle or related 1,3 interatomic distance can be determined from the diffraction data is poor compared with that of the predicted geometry. Angle restraints should be worthwhile in structures at resolution $\gtrsim$2.0 Å. Angle restraints will normally be effected by restraining the distance between two donor atoms ($L_1\cdots L_2$) with a 'DFIX' type of instruction (*SHELXL*; Sheldrick & Schneider, 1997). Table 6 gives a very simple set of numerical values for doing this and Table S2 gives fuller details of how they have been established. Many simplifications were made, which were justified by examining data for complexes in the CSD (and PDB, resolution ≤1.6 Å). For example, distributions of angles such as N—M—N, N—M—O and O—M—O are very similar, as are the distributions of N···N, N···O and O···O distances in complexes. Thus, donor atoms are simplified to two types, N,O (may be N or O) and S (whose parameters could also be used for Cl). M—S distances were treated as equal to 1.10 times the equivalent M—N,O distance (see Table S2). The standard deviations suggested are based on standard deviations of 0.10 Å for the M—L distance, 4° in the L—M—L angle for octahedral coordination and 8° in the L—M—L angle for tetrahedral coordination. The distributions of $L\cdots L$ distances in unchelated metal complexes in the CSD were examined and three are illustrated in Fig. 4. Fig. 4(*a*) is a well behaved example, Mn with coordination number 6, where both distance and standard deviation from Table 6 fit well. In Fig. 4(*b*), Zn with coordination number 4, the range of observed distances is, as expected, much wider, but the observed mean distance does not fit well with the distance given in Table 6, perhaps because of the different chemical nature of ligands or differences in Zn—N,O distances. Lastly, Fig. 4(*c*) illustrates the distribution for Cu^{II}, coordination number 6, and the difficulties in devising restraints in this situation; owing to the Jahn–Teller effect, two types of $L\cdots L$ distance are present. For validation of all these it will be appropriate to use the restraint targets given, ±3 standard deviations.

An alternative type of restraint may be considered, described as 'BUMP' in *SHELXL* (Sheldrick & Schneider, 1997). It simply prevents atoms approaching closer than a van der Waals contact distance or other specified distance (unless they are linked through covalent bonds other than those to the metal, *e.g.* the two O atoms of a carboxylate group). This would be applied to all pairs of donor atoms $L1$, $L2$ in the first coordination sphere of the metal which are not otherwise linked through one, two, three or four covalent bonds (as are, for example, the two O atoms of a carboxylate group). This is a physically realistic restraint, which could be both simple and quite useful for five-, six-, seven- (and probably eight-) coordinate groups, assuming that all the donor atoms have been identified. It would not be effective in keeping a four-coordinate group tetrahedral. It is not incompatible with the DFIX type of restraint above. Parameters devised from the CSD are given in Table 6(*b*), with fuller information in Table S3. The values found for N,O$\cdots$N,O are longer than the default distances provided by *SHELXL*, which allows for the presence of hydrogen bonding between the atom pairs; this would not normally be present within a metal coordination group. The values for N,O$\cdots$N,O do correspond closely to the sum of van der Waals radii as given, for example, by Bondi (1964). However, for N,O$\cdots$S,Cl and S,Cl$\cdots$S,Cl, Bondi's radii give 3.3 and 3.6 Å, significantly longer than the values found here in ML_5, ML_6 or square-planar ML_4 complexes. In these complexes, adjacent S or Cl atoms are forced closer together than the preferred van der Waals contact distance, thus making these geometries less favourable than tetrahedral ML_4 when several of the ligands have S or Cl donor atoms.

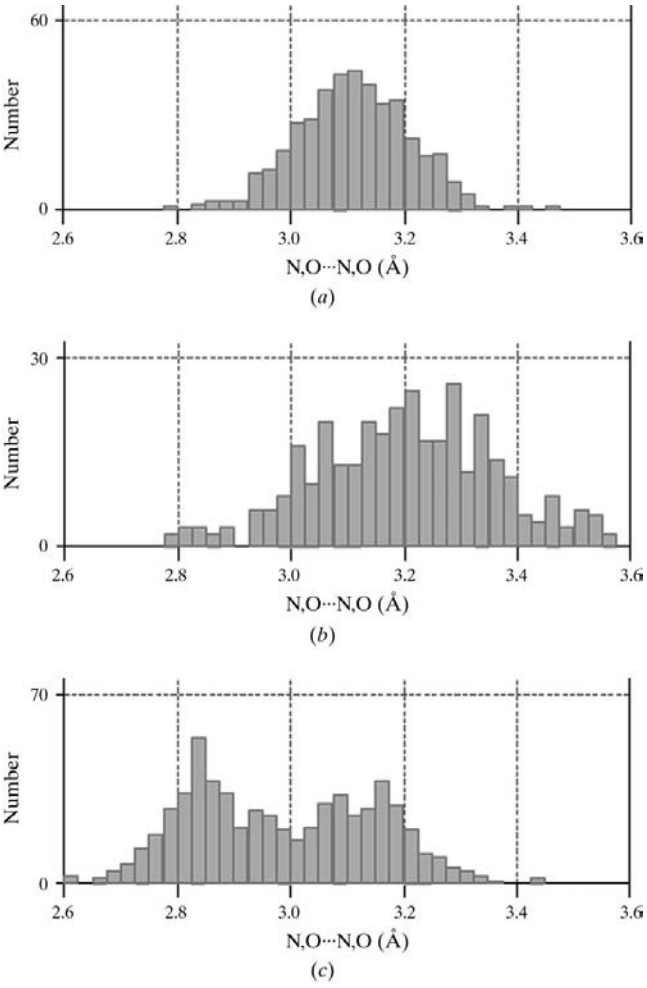


Figure 4
Examples of distributions of N,O$\cdots$N,O non-bonded distances (in Å) in non-chelated metal complexes in the CSD, as a guide to the possible efficacy of restraints on angles. (*a*) Mn with coordination number 6, (*b*) Zn with coordination number 4, (*c*) Cu^{II} with coordination number 6.

Table 6
Donor$\cdots$donor distances, $L_1\cdots L_2$ (Å), for restraints or validation of angles at metal atoms.

(*a*) $L_1\cdots L_2$ to give ideal octahedral or tetrahedral geometry when predicted $M-L$ distances are used and standard deviations. The standard deviations are large because of the flexibility of the angles and the additional uncertainties in predicting metal–donor distances. For this reason, only a very simple system is proposed (for fuller details and justification, see supplementary Table S2). In octahedral ML_6 only 'adjacent' ligands in the coordination sphere are restrained (*i.e.* those where the angle is ~180°, $L\cdots L > 3.8$ Å are excluded).

Distances are for $L_1 = L_2$ = N,O. For L_1 = N,O, L_2 = S, distances should be increased by a factor of 1.05; for $L_1 = L_2$ = S, distances should be increased by a factor of 1.10.

	Ca	Mg	Mn	Fe	Cu^{I}	Zn	Standard deviation
Coordination number 6, octahedral	3.38	3.08	3.15	2.91	—	2.90	0.15
Coordination number 4, tetrahedral	—	3.56	3.64	3.36	3.56	3.35	0.20

(*b*) $L_1\cdots L_2$ for closest allowed approach of donor atoms in non-chelated metal coordination groups in the CSD. These may be used in a 'BUMP' type restraint (*SHELXL*; Sheldrick & Schneider, 1997) and should be useful for coordination numbers 5, 6, 7 (and probably 8), but not 4. Table S3 gives fuller details. Use of the distances and standard deviations given here will give a distribution of $L_1\cdots L_2$ distances similar to those found in metal complexes in the CSD.

N,O$\cdots$N,O	3.00 (15)
N,O$\cdots$S,Cl	3.19 (12)
S,Cl$\cdots$S,Cl	3.38 (9)

4. Conclusions

There is plenty of evidence to indicate that these metals show the same geometrical behaviour in proteins and in relevant small-molecule complexes. Except for the highest resolution protein structure determinations (better than 1.6 Å), the precision of bond lengths and angles determined in proteins is not as good as the precision with which they can be predicted and it should be worthwhile to apply restraints in refinement. Target distances and standard deviations for such restraints are provided (Tables 1 and 6), as well as a summary of preferred coordination numbers and shapes for the complexes of each metal (Table 4). These are also appropriate for validation and the survey suggests that significant deviations from these target distances should be examined – if they are not the result of errors in interpretation or refinement, they could have real biochemical significance.

Metals other than Ca, Mg, Mn, Fe, Cu and Zn are not often found in proteins in their natural state, but for a selection of them and for Na^+ and K^+ some further work is planned to establish suitable target distances. Also, recent development of the CSD has allowed more reliable identification of the oxidation states of metals such as Mn, Fe and Cu (Shields *et al.*, 2000) and thus evaluation of separate target distances for Fe^{II}, Fe^{III} *etc.* has become practicable and would be worthwhile to perform.

Within the University of Edinburgh, I am very grateful to Professor Malcolm Walkinshaw for access to computing facilities, to Dr Paul Taylor for computational support and to both for helpful discussions. I am also grateful to Alex Gruneau and Jenny Arscott who did preliminary explorations of these topics. I also thank Dr Greg Shields and others at the Cambridge Crystallographic Data Centre and Dr Kim Henrick at the European Bioinformatics Institute (Hinxton, Cambridge) for advice and comments.

References

Allen, F. H. & Kennard, O. (1993*a*). *Chem. Des. Autom. News*, **8**, 1.

Allen, F. H. & Kennard, O. (1993*b*). *Chem. Des. Autom. News*, **8**, 31–37.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *Nucleic Acids Res.* **28**, 235–242.

Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. E., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *J. Mol. Biol.* **112**, 535.

Bondi, A. (1964). *J. Phys. Chem.* **68**, 441–451.

Chakrabarti, P. (1990*a*). *Biochemistry*, **29**, 651–658.

Chakrabarti, P. (1990*b*). *Protein Eng.* **4**, 49–56.

Chakrabarti, P. (1990*c*). *Protein Eng.* **4**, 57–63.

Chong, K. T., Miyazaki, G., Morimoto, H., Oda, Y. & Park, S.-Y. (1999). *Acta Cryst.* D**55**, 1291–1300.

Cruickshank, D. W. J. (1999). *Acta Cryst.* D**55**, 583–601.

Engh, R. A. & Huber, R. (1991). *Acta Cryst.* A**47**, 392–400.

Frausto da Silva, J. J. R. & Williams, R. J. P. (1991). *The Biological Chemistry of the Elements.* Oxford: Clarendon Press.

Harding, M. M. (1999). *Acta Cryst.* D**55**, 1432–1443.

Harding, M. M. (2000). *Acta Cryst.* D**56**, 857–867.

Maher, M. J., Xiao, Z., Wilce, M. C. J., Guss, J. M. & Wedd, A. G. (1999). *Acta Cryst.* D**55**, 962–968.

Parisini, E., Capozzi, F., Lubini, P., Lamzin, V., Luchinat, C. & Sheldrick, G. M. (1999). *Acta Cryst.* D**55**, 1773–1784.

RCSB (1999). *Research Collaboratory for Structural Bioinformatics, Protein Data Bank, Release #89, July 1999.* Rutgers, The State University of New Jersey, New Brunswick, New Jersey, USA.

Sheldrick, G. M. & Schneider, T. R. (1997). *Methods Enzymol.* **277**, 319–343.

Shields, G. P., Raithby, P. R., Allen, F. H. & Motherwell, W. D. S. (2000). *Acta Cryst.* B**56**, 455–465.

Straeter, N. & Lipscomb, W. N. (1995). *Biochemistry*, **34**, 14792–14800.