FindGeo: a tool for determining metal coordination geometry

SUPPLEMENTARY MATERIAL

Claudia Andreini[1, 2, *], Gabriele Cavallaro[1], and Serena Lorenzini[1]

[1]Magnetic Resonance Center (CERM) – University of Florence, Via L. Sacconi 6, 50019 Sesto Fiorentino, Italy

[2]Department of Chemistry – University of Florence, Via della Lastruccia 3, 50019 Sesto Fiorentino, Italy

Appendix A

RMSD-based definition of the criteria for determining coordination geometry

According to the most recent IUPAC recommendations for the nomenclature of inorganic chemistry contained in the so-called “Red Book” (Connelly and others 2005), the description of the geometrical configuration of a coordination compound “should be based on the nearest idealized geometry”. It is also noted, however, that “care may be required in making the choice” as some idealized geometries are closely related. The algorithm implemented in FindGeo directly addresses both of these recommendations by using RMSD as a metric for evaluating the similarity between the overall arrangement of the coordinating atoms around a metal and various possible idealized geometries. In fact, the ranking of the alternative geometries based on their RMSD values allows not only the straightforward identification of the closest idealized geometry, but also the quantitative comparison of the possible geometry assignments. The RMSD values yielded by FindGeo can be viewed as the distances between the actual geometry of the input site and the individual idealized geometries, reducing the complex spatial relationships among three-dimensional objects to single numbers. In this view, we used the structural templates with idealized geometries as inputs for FindGeo, so as to obtain a set of matrices containing the pairwise distances (i.e., the RMSD values) between all the idealized geometries defined for a given coordination number. These matrices (shown in Table 1) are symmetric, indicating that FindGeo calculates the same distance between two geometries regardless of which one is used as input. While the numerical values in these matrices would change if a different value (i.e., not 3 Å) had been chosen for the metal-coordinating atom distances, the distance relationships among the different idealized geometries they describe are of general validity: for example, it can be stated that the closest idealized geometries for coordination number four are *bvp* and *pyv*, while those for coordination number six are *oct* and *ctn* (Table 1).

We used the distance matrices described above to define two threshold RMSD values for each idealized geometry, which we called the *assignment threshold* and the *distortion threshold*.

These thresholds determine whether the idealized geometry with the lowest RMSD is indeed a reasonable assignment for the input structure by taking into account the absolute value of that RMSD, and are used to provide a qualitative (and thus more immediately understood by the user) assessment of the geometry assignment produced. Specifically, the assignment threshold is the RMSD value beyond which an idealized geometry, despite being the nearest match for the input structure, is not considered a valid assignment, while the distortion threshold is the RMSD value beyond which an idealized geometry is still considered a valid assignment but is labelled as distorted. For a given idealized geometry, the assignment threshold is defined as half of the maximum RMSD value calculated for that geometry (i.e., half of the largest number in the corresponding matrix row), and the distortion threshold is defined as half of the minimum RMSD value calculated for that geometry (i.e., half of the smallest number in the corresponding matrix row).

The rationale for the definition of the above thresholds can be exemplified by the case of coordination number 3, as depicted in Figure 1. In this Figure, the three possible idealized geometries (*tri*, *tev*, and *spv*) are represented as points on a plane whose pairwise distances correspond to the RMSD values shown in Table 1. The lines joining two idealized geometries represent the continuous set of non-idealized geometries that are intermediate between the two, and that can be thought to be generated by progressively distorting one geometry until the other is reached. For example, if the *tri* geometry is distorted towards the *tev* geometry (pathway 1 in Figure 1) the value of the RMSD from *tri* can increase up to 1.015/2 Å (corresponding to the midpoint of the pathway) before the best geometry assignment becomes *tev* (i.e., the RMSD from *tev* becomes lower than that from *tri*). If instead the *tri* geometry is distorted towards the *spv* geometry (pathway 2 in Figure 1) the value of the RMSD from *tri* can increase up to 1.268/2 Å before the best geometry assignment becomes *spv*. Finally, if the *tri* geometry is distorted in a way that does not lead to either *tev* or *spv* (pathway 3 in Figure 1) the best geometry assignment remains *tri* whatever the RMSD value. This latter type of distortion could produce meaningless results if the absolute value of the RMSD was not considered, as FindGeo would classify as *tri* even structures that are so remotely related to the idealized *tri* geometry that they could hardly be recognized as such by visual inspection. This problem is however avoided by the introduction of the two above described thresholds, which in the representation of Figure 1 define two concentric circles centred on each idealized geometry. For the *tri* geometry, the inner circle (with solid border in Figure 1) has a radius equal to the distortion threshold (set to 1.015/2 Å for *tri*) and delimits all the geometries that can be readily assigned as *tri*, as within this circle no other idealized geometry can be closer to the geometry of the input structure than *tri*. This circle can thus be thought of as the “core” region of existence of the *tri* geometry. The outer circle (with dashed border in Figure 1) has a radius equal to the assignment threshold (set to 1.268/2 Å for *tri*) and, compared with the inner circle, delimits additional geometries for which *tri* may (e.g., for a geometry along pathway 2) or may not (e.g., for a geometry along pathway 1) be the closest to the geometry of the input structure. In the former case, the geometry of the input structure is described as distorted *tri*, as it is at the same distance from the idealized *tri* geometry as geometries which are closer to a different idealized geometry (e.g., *tev* for a geometry along pathway 1). The outer circle can thus be thought of as the “extended” region of existence of the *tri* geometry, with the annulus between the outer and the inner circle being the region of existence of the distorted *tri* geometry. Outside the outer circle, the geometry can no longer be described as *tri* even if *tri* is the closest idealized geometry (e.g., for a geometry along pathway 3), as it is at a distance from the idealized *tri* geometry at which any other idealized geometry becomes possible. In this case, the geometry of the input structure is described as irregular.

It is important to note that when the idealized geometry that is ranked first by FindGeo (i.e., that with the lowest RMSD from the input structure) has an RMSD value beyond its assignment threshold, other idealized geometries may have higher RMSD values but below their assignment threshold. In these cases, the idealized geometry with the lowest RMSD among those below their assignment threshold is selected as the best estimate of the geometry, although it is not the closest one to the input structure. When this occurs, FindGeo displays a warning message advising that the best estimate is, in a way, a suboptimal solution, and additionally reports the idealized geometry that was discarded despite having the lowest RMSD value.

Appendix B

Application of FindGeo to test data sets: artificial data set

This data set was generated starting from the structural templates with idealized geometries included in the library of FindGeo, by applying random perturbations to the positions of the coordinating atoms. In detail, for every idealized geometry we first generated three sets of 500 structures each in which the metal-coordinating atom bonds were tilted from their original positions and randomly relocated within a cone around the original bond axis. The maximum allowed tilt angle was set to 2.5°, 5° and 10° for the three sets, respectively. We then retained only the structures that had all coordinating atom-metal-coordinating atom (CA-M-CA hereafter) angles within 2.5°, 5° and 10° (respectively for the three sets) from their values in the corresponding idealized geometries. This filtering was done to match the criteria used to construct the CSD-derived data set (see below). Finally, we selected from each set the 100 structures whose CA-M-CA angles had the largest average deviation from their values in the corresponding idealized geometries. This procedure, which was implemented in a separate program making use of the p3d Python module (Fufezan,C. and Specht,M. 2009) to measure CA-M-CA angles, thus resulted in having 300 structures for each idealized geometry, out of which 100 can be described as slightly distorted (those obtained by allowing 2.5° tilt angles), 100 moderately distorted (5° tilt angles), and 100 strongly distorted (10° tilt angles).

A summary of the results obtained for this data set is reported in Table 2. For all the 10800 structures but one, the idealized geometry from which the distorted geometry was derived was identified as the best possible assignment. In the single case of a strongly distorted *csa* geometry FindGeo classified it as irregular, based on an RMSD value (0.294 Å, see Table 2) slightly above the assignment threshold for this geometry (0.292 Å, see Table 1). In two more cases, a strongly distorted *bvp* geometry and a strongly distorted *pyv* geometry were classified as distorted, as their RMSD values (0.278 Å and 0.288 Å respectively, see Table 2) were above the respective distortion thresholds (0.277 Å for both, see Table 1) but below the respective assignment thresholds (0.650 Å and 0.812 Å respectively, see Table 1).

A more detailed view of the results summarized in Table 2 is provided in Supplementary Figure S2 as a series of plots, each corresponding to one of the sets of 100 structures described above and showing the RMSD values calculated for every geometry tested on those structures. An example plot is shown in Figure 2A, which illustrates the results obtained for the 100 structures with strongly distorted *oct* geometries. Figure 2A indicates that the distortion of the *oct* geometry determined by 10° tilt angles resulted in RMSD values for this geometry ranging between approximately 0.17 and 0.25 Å. The RMSD values for the other geometries were well above this range, although they were in all cases lower than those calculated using the idealized *oct* geometry as input (dashed lines in Figure 2A).

Application of FindGeo to test data sets: CSD-derived data set

This data set was generated by searching the CSD for high-resolution X-ray structures of complexes of 12 biologically important metals (including Ca, Co, Cu, Fe, K, Mg, Mn, Mo, Na, Ni, V, and Zn), using the CSD interface program ConQuest (Bruno,I.J. *et al.* 2002) to retrieve complexes with specific coordination geometries. In detail, for every idealized geometry included in the library of FindGeo we built a query specifying the corresponding coordination number and the values of all the possible CA-M-CA angles, which were required to be within a specified tolerance from their ideal values. For coordination numbers higher than four, however, constraining all the possible CA-M-CA angles (whose number is equal to $n(n - 1)/2$ being n the coordination number) made the CSD search to become slow and practically unfeasible. In these cases, therefore, only a subset (i.e., six) of the CA-M-CA angles were constrained in the query, and the structures returned by the search were subsequently filtered based on whether the values of the other CA-M-CA angles (i.e., those excluded from the query) were also acceptable (i.e., within the specified tolerance) or not. By analogy to the procedure described above for the artificial data set, we used three different tolerances (2.5°, 5° and 10°), thereby obtaining for each geometry three sets of structures which we refer to as slightly distorted (those retrieved using a 2.5° tolerance), moderately distorted (5° tolerance), and strongly distorted (10° tolerance).

Overall, this data set included 3264 slightly distorted, 5235 moderately distorted, and 5843 strongly distorted structures. The CSD reference codes of these structures are listed in Supplementary Table S2. Not unexpectedly, only some common coordination geometries were represented in this data set, with the octahedral (*oct*) and the square planar (*spl*) being by far the most frequent geometries (8009 and 3554 structures, respectively). A summary of the results obtained for this data set is reported in Table 3. For almost all structures, the geometry defined on the basis of the CA-M-CA angles was identified as the best geometry assignment. The only two exceptions were a moderately distorted and a strongly distorted *tri* structures, which were both classified as *tev*. The former is the structure of the $NiCu(CO)_2(\mu\text{-}Ph_2PCH_2PPh_2)_2(BH_3CN)$ complex (CSD reference code POGBES) (Holah,D.G. *et al.* 1994), in which a Cu(II) ion is coordinated by one nitrogen and two phosphorus atoms, with CA-M-CA angles of 116.2°, 116.8° and 117.6° and the Cu(II) ion lying approximately 0.38 Å above the plane identified by the three coordinating atoms. In this case, the RMSD calculated for the *tri* geometry was 0.536 Å (i.e., above the distortion threshold but below the assignment threshold for this geometry, see Table 1), and that calculated for the *tev* geometry was 0.483 Å (i.e., below the distortion threshold for this geometry, see Table 1). The latter is the structure of the $[Cu\{\kappa^3\text{-}H(\mu\text{-}H)B(tiaz)_{2\}}(PPh_3)]$ complex (tiaz = 2-mercaptothiazolyl, CSD reference code NOLMOR) (Beheshti,A. *et al.* 2008), in which a Cu(I) ion is coordinated by one phosphorus and two sulfur atoms, with CA-M-CA angles of 111.7°, 117.7° and 121.6° and the Cu(I) ion lying approximately 0.39 Å above the plane identified by the three coordinating atoms. In this case, the RMSD calculated for the *tri* geometry was 0.539 Å (i.e., above the distortion threshold but below the assignment threshold for this geometry, see Table 1), and that calculated for the *tev* geometry was 0.508 Å (i.e., equal to the distortion threshold for this geometry, see Table 1).

A more detailed view of the results summarized in Table 3 is provided in Supplementary Figure S3 as a series of plots analogous to those presented for the artificial data set (Supplementary Figure S2). For the sake of comparison, an example plot illustrating the results obtained for the 3144 structures with strongly distorted *oct* geometries is shown in Figure 2B. This plot substantially resembles that obtained for the artificially distorted *oct* geometries (Figure 2A), except that the much higher number of structures is reflected in a wider range of RMSD values for all geometries. Still, the range of RMSD values calculated for the *oct* geometry (between approximately 0.11 and 0.35 Å) was again well separated from those calculated for the other geometries.

Application of FindGeo to test data sets: metalloprotein data set

This data set consisted of the metal sites in metalloproteins that were examined for coordination geometry in a previous work based on the analysis of CA-M-CA angles (Rulisek,L. and Vondrasek,J. 1998). It was generated by using the feature of FindGeo that allows the automatic download and identification of metal sites in PDB structures, starting from the list of PDB accession codes reported in (Rulisek and Vondrasek 1998). After discarding 18 sites whose reported geometry was either not specified (e.g., a dicobalt site in the PDB structure 1mat was described as "$(Co)_2$ bridged") or inconsistent with the reported ligands (e.g., one of the two zinc sites in the PDB structure 1hra was described as "trigonal bipyramidal" even if it has four Cys ligands), the data set included a total of 136 sites (18 Cd, 9 Co, 30 Cu, 14 Hg, 10 Ni, and 55 Zn) found in 106 metalloproteins. In 113 cases, the reported ligands were retrieved by leaving the default threshold value of 2.8 Å for the coordination distance, whereas in the other 23 cases it was necessary to use a higher (10 cases, up to 3.3 Å) or a lower (13 cases, down to 2.5 Å) threshold.

A summary of the results obtained for this data set is given in Supplementary Table S3. In 101 out of 136 cases (i.e., about 74%), the best geometry assignment was in agreement with the assignment reported in (Rulisek and Vondrasek 1998). Figure 3 (panel A) shows a histogram of the specific geometry assignments for these 101 sites. These include 15 sites (14%) in which the geometry was one with a vacancy, 28 sites (26%) in which the geometry was described by FindGeo as distorted, and 4 sites (4%) in which the geometry was described by FindGeo (as well as in (Rulisek and Vondrasek 1998)) as irregular. In three cases (a copper site in the PDB structure 1lfi and two cobalt sites in the PDB structures 1cah and 1lna, respectively), the geometry with the lowest RMSD from the input structure was discarded because that RMSD value was beyond the assignment threshold, and the best geometry resulted to be distorted octahedral in all the three cases.

In the remaining 35 cases (i.e., about 26%), the best geometry indicated by FindGeo did not match that reported in (Rulisek and Vondrasek 1998). Figure 3 (panel B) shows a histogram of the specific geometry assignments for these 35 sites. As detailed in Table 4, three types of mismatch occurred: (i) in 24 cases (68%) FindGeo identified a geometry different from that reported in (Rulisek and Vondrasek 1998); in particular, in 22 of these cases FindGeo identified a geometry with a vacancy as the best one, whereas the reported geometry was one with no empty coordination positions; (ii) in 9 cases (26%) FindGeo classified the geometry as irregular, whereas a geometry could be identified in (Rulisek and Vondrasek 1998); in 6 of these cases, the geometry with the lowest RMSD value coincided with the one reported in (Rulisek and Vondrasek 1998); (iii) in 2 cases (6%) FindGeo was able to identify a geometry whereas the geometry was classified as irregular in (Rulisek and Vondrasek 1998).

Altogether, FindGeo classified 44 geometries (32%) as distorted, and 13 (10%) as irregular. The degree of mismatch between the results of FindGeo and the geometries reported in (Rulisek and Vondrasek 1998) increased when going from regular to distorted to irregular geometries (Figure 4, panel A). Also, the degree of mismatch varied with the metal present in the site (being highest for copper and nickel, Figure 4, panel B) and with the coordination number (being highest for five, Figure 4, panel C).

Appendix C

A note on the protonation status of metal ligands

The vast majority of X-ray structures are determined at resolutions that do not allow hydrogen atoms to be observed. The assignment of hydrogen coordinates to protein structures is a far from trivial task, which is typically approached by complex quantum chemical calculations (Nilsson,K. and Ryde,U. 2004; Ryde,U. 2007; Labute,P. 2009). Clearly, this task is not within the scope of FindGeo, which exclusively uses the information contained in the coordinates of the input structure to determine metal coordination geometry. In practice, there is no need to determine the protonation status of (potential) metal ligands prior to FindGeo calculations. In fact, if ligand protonation affected metal ligation, the ligand would move well beyond typical coordination distances, as shown by quantum chemical calculations (see, e.g., (Su,P. and Li,H. 2010)). This observation can be interpreted as resulting from the mutually exclusive competition between the metal and the proton for the same electron pair. Therefore, if a ligand is found in the structure at coordination distance from the metal (which is the criterion used by FindGeo to determine metal ligands) this can be generally taken as the guarantee that the donor atom is either not protonated or that protonation does not interfere with metal ligation (i.e., another different electron pair is involved in binding the proton).

References

Beheshti,A. et al. (2008) Complexes of copper(I) and silver(I) with bis(methimazolyl)borate and dihydrobis(2-mercaptothiazolyl)borate ligands. *Dalton Trans*,6641-6646.

Bruno,I.J. et al. (2002) New software for searching the Cambridge Structural Database and visualizing crystal structures. *Acta Crystallogr B*, 58, 389-397.

Connelly NG, Damhus T, Hartshorn RM, Hutton AT. 2005. Nomenclature of Inorganic Chemistry - IUPAC Recommendations 2005. RSC Publishing.

Fufezan,C. and Specht,M. (2009) p3d--Python module for structural bioinformatics. *BMC Bioinformatics*, 10, 258.

Holah,D.G. et al. (1994) New bis(diphenylphosphino)methane-bridged d^{10}-d^{10} heterobinuclear complexes derived from $Ni(Co)_2(\eta^1\text{-}Ph_2PCH_2PPh_2)_2$ and $Ni_2(\mu\text{-}CO)(CO)_2(\mu\text{-}Ph_2PCH_2PPh_2)_2$: The crystal structure of $NiCu(CO)_2(\mu\text{-}Ph_2PCH_2PPh_2)_2(BH_3CN)$. *Polyhedron*, 13, 2431-2437.

Labute,P. (2009) Protonate3D: assignment of ionization states and hydrogen coordinates to macromolecular structures. *Proteins*, 75, 187-205.

Nilsson,K. and Ryde,U. (2004) Protonation status of metal-bound ligands can be determined by quantum refinement. *J Inorg Biochem*, 98, 1539-1546.

Rulisek,L. and Vondrasek,J. (1998) Coordination geometries of selected transition metal ions (Co2+, Ni2+, Cu2+, Zn2+, Cd2+, and Hg2+) in metalloproteins. *J Inorg Biochem*, 71, 115-127.

Ryde,U. (2007) Accurate metal-site structures in proteins obtained by combining experimental data and quantum chemistry. *Dalton Trans*,607-625.

Su,P. and Li,H. (2010) Protonation of type-1 Cu bound histidines: a quantum chemical study. *Inorg Chem*, 49, 435-444.

Table 1. Matrices containing the RMSD values between the idealized geometries included in FindGeo, calculated by using the corresponding ideal site structures as input (CN = coordination number; CG = coordination geometry; DT = distortion threshold; AT = assignment threshold; see text for details and Supplementary Table S1 and Supplementary Figure S1 for the names of the geometries). All the values are in Å.

CN	CG									DT	AT
2		**Lin**	**Trv**								
	Lin		1.553							0.777	0.777
	Trv	1.553								0.777	0.777
3		**Tri**	**Tev**	**Spv**							
	Tri		1.015	1.268						0.508	0.634
	Tev	1.015		1.484						0.508	0.742
	Spv	1.268	1.484							0.634	0.742
4		**Tet**	**Spl**	**Bva**	**Bvp**	**Pyv**					
	Tet		1.817	0.879	1.300	1.335				0.440	0.909
	Spl	1.817		1.960	1.098	1.624				0.549	0.980
	Bva	0.879	1.960		1.230	1.098				0.440	0.980
	Bvp	1.300	1.098	1.230		0.554				0.277	0.650
	Pyv	1.335	1.624	1.098	0.554					0.277	0.812
5		**Tbp**	**Spy**	**Tpv**							
	Tbp		0.982	1.101						0.491	0.551
	Spy	0.982		1.207						0.491	0.604
	Tpv	1.101	1.207							0.551	0.604
6		**Oct**	**Tpr**	**Pva**	**Pvp**	**Cof**	**Con**	**Ctf**	**Ctn**		
	Oct		1.255	1.780	0.853	0.996	1.673	1.291	1.202	0.427	0.890
	Tpr	1.255		1.341	1.026	1.202	1.175	0.997	1.496	0.499	0.748
	Pva	1.780	1.341		1.490	1.064	1.064	0.858	1.414	0.429	0.890
	Pvp	0.853	1.026	1.490		0.863	1.049	1.007	0.855	0.427	0.745
	Cof	0.996	1.202	1.064	0.863		1.214	0.970	0.977	0.432	0.607
	Con	1.673	1.175	1.064	1.049	1.214		1.136	1.228	0.525	0.837
	Ctf	1.291	0.997	0.858	1.007	0.970	1.136		0.915	0.429	0.646
	Ctn	1.202	1.496	1.414	0.855	0.977	1.228	0.915		0.428	0.748
7		**Pbp**	**Coc**	**Ctp**	**Hva**	**Hvp**	**Cuv**	**Sav**			
	Pbp		0.999	0.950	1.651	0.746	1.098	0.960		0.373	0.826
	Coc	0.999		1.115	1.567	0.849	0.763	1.010		0.382	0.784
	Ctp	0.950	1.115		1.537	1.154	1.132	0.651		0.326	0.769
	Hva	1.651	1.567	1.537		1.468	0.939	1.309		0.470	0.826
	Hvp	0.746	0.849	1.154	1.468		0.836	1.116		0.373	0.734
	Cuv	1.098	0.763	1.132	0.939	0.836		0.939		0.382	0.566
	Sav	0.960	1.010	0.651	1.309	1.116	0.939			0.326	0.655
8		**Hbp**	**Cub**	**Sqa**	**Boc**	**Bts**	**Btt**				
	Hbp		0.879	1.283	0.916	1.338	1.513			0.440	0.757
	Cub	0.879		0.956	0.714	1.147	1.491			0.357	0.746
	Sqa	1.283	0.956		1.182	0.618	1.594			0.309	0.797
	Boc	0.916	0.714	1.182		1.262	1.087			0.357	0.631
	Bts	1.338	1.147	0.618	1.262		1.540			0.309	0.770
	Btt	1.513	1.491	1.594	1.087	1.540				0.544	0.797
9		**Ttp**	**Csa**								
	Ttp		0.583							0.292	0.292
	Csa	0.583								0.292	0.292

Table 2. Results of FindGeo on the artificial data set, consisting of three sets of 100 structures (shown under 2.5°, 5° and 10° tilt angles, respectively) for each idealized geometry (CN = coordination number; CG in = idealized geometry from which the distorted geometries were derived; CG out = geometries assigned by FindGeo and corresponding number of instances; RMSD range = range of RMSD values calculated for each geometry assigned by FindGeo, in Å; see text for details and Supplementary Table S1 and Supplementary Figure S1 for the names of the geometries).

		2.5° tilt angles		**5.0° tilt angles**		**10.0° tilt angles**	
CN	**CG in**	**CG out**	**RSMD range**	**CG out**	**RMSD range**	**CG out**	**RMSD range**
2	**Lin**	100 lin	0.050-0.065	100 lin	0.103-0.130	100 lin	0.196-0.262
	Trv	100 trv	0.040-0.065	100 trv	0.078-0.131	100 trv	0.133-0.261
3	**Tri**	100 tri	0.043-0.079	100 tri	0.084-0.173	100 tri	0.150-0.387
	Tev	100 tev	0.040-0.070	100 tev	0.085-0.143	100 tev	0.158-0.338
	Spv	100 spv	0.044-0.070	100 spv	0.090-0.146	100 spv	0.161-0.301
4	**Tet**	100 tet	0.044-0.071	100 tet	0.089-0.150	100 tet	0.169-0.306
	Spl	100 spl	0.042-0.065	100 spl	0.082-0.135	100 spl	0.144-0.265
	Bva	100 bva	0.043-0.083	100 bva	0.092-0.155	100 bva	0.183-0.313
	Bvp	100 bvp	0.043-0.061	100 bvp	0.079-0.126	99 bvp 1 bvp distorted	0.161-0.266 0.278
	Pyv	100 pyv	0.047-0.068	100 pyv	0.094-0.131	99 pyv 1 pyv distorted	0.182-0.275 0.288
5	**Tbp**	100 tbp	0.042-0.071	100 tbp	0.085-0.160	100 tbp	0.184-0.313
	Spy	100 spy	0.045-0.065	100 spy	0.088-0.126	100 spy	0.181-0.278
	Tpv	100 tpv	0.043-0.075	100 tpv	0.087-0.144	100 tpv	0.175-0.298
6	**Oct**	100 oct	0.042-0.070	100 oct	0.084-0.130	100 oct	0.170-0.246
	Tpr	100 tpr	0.042-0.063	100 tpr	0.087-0.147	100 tpr	0.185-0.261
	Pva	100 pva	0.040-0.069	100 pva	0.085-0.155	100 pva	0.177-0.340
	Pvp	100 pvp	0.042-0.067	100 pvp	0.086-0.148	100 pvp	0.180-0.335
	Cof	100 cof	0.042-0.071	100 cof	0.081-0.134	100 cof	0.167-0.242
	Con	100 con	0.044-0.067	100 con	0.086-0.137	100 con	0.182-0.269
	Ctf	100 ctf	0.046-0.078	100 ctf	0.085-0.153	100 ctf	0.193-0.323
	Ctn	100 ctn	0.044-0.074	100 ctn	0.090-0.153	100 ctn	0.178-0.302
7	**Pbp**	100 pbp	0.041-0.071	100 pbp	0.084-0.143	100 pbp	0.185-0.337
	Coc	100 coc	0.040-0.072	100 coc	0.082-0.135	100 coc	0.168-0.252
	Ctp	100 ctp	0.044-0.077	100 ctp	0.086-0.154	100 ctp	0.190-0.310
	Hva	100 hva	0.042-0.063	100 hva	0.083-0.134	100 hva	0.176-0.333
	Hvp	100 hvp	0.041-0.061	100 hvp	0.083-0.126	100 hvp	0.169-0.269
	Cuv	100 cuv	0.044-0.062	100 cuv	0.083-0.126	100 cuv	0.180-0.263
	Sav	100 sav	0.046-0.062	100 sav	0.089-0.125	100 sav	0.179-0.288
8	**Hbp**	100 hbp	0.040-0.060	100 hbp	0.084-0.126	100 hbp	0.173-0.273
	Cub	100 cub	0.042-0.061	100 cub	0.084-0.117	100 cub	0.175-0.259
	Sqa	100 sqa	0.043-0.060	100 sqa	0.087-0.131	100 sqa	0.171-0.249
	Boc	100 boc	0.038-0.071	100 boc	0.075-0.134	100 boc	0.144-0.238
	Bts	100 bts	0.043-0.074	100 bts	0.087-0.146	100 bts	0.199-0.292
	Btt	100 btt	0.039-0.062	100 btt	0.078-0.128	100 btt	0.161-0.265
9	**Ttp**	100 ttp	0.045-0.069	100 ttp	0.089-0.136	100 ttp	0.196-0.289
	Csa	100 csa	0.042-0.060	100 csa	0.086-0.146	99 csa 1 irregular	0.181-0.280 0.294

Table 3. Results of FindGeo on the CSD-derived data set, consisting of three sets of structures (shown under 2.5°, 5° and 10° tolerance, respectively) for each idealized geometry (CN = coordination number; CG in = idealized geometry from which the values of the L-M-L angles were derived to query the CSD; NS = number of structures retrieved in the CSD; CG out = geometries assigned by FindGeo and corresponding number of instances; RMSD range = range of RMSD values calculated for each geometry assigned by FindGeo, in Å; see text for details and Supplementary Table S1 and Supplementary Figure S1 for the names of the geometries).

		2.5° tolerance			**5° tolerance**			**10° tolerance**		
CN	**CG in**	**NS**	**CG out**	**RMSD range**	**NS**	**CG out**	**RMSD range**	**NS**	**CG out**	**RMSD range**
2	**Lin**	81	81 lin	0.000-0.066	46	46 lin	0.066-0.127	48	48 lin	0.130-0.251
3	**Tri**	66	66 tri	0.001-0.462	66	65 tri 1 tev	0.060-0.334 0.483[1]	90	89 tri 1 tev	0.113-0.476 0.508[2]
4	**Tet**	214	214 tet	0.000-0.108	483	483 tet	0.051-0.216	883	883 tet	0.100-0.416
	Spl	1324	1324 spl	0.000-0.100	1184	1184 spl	0.059-0.191	1046	1046 spl	0.124-0.356
5	**Tpb**	26	26 tbp	0.000-0.062	101	101 tbp	0.047-0.180	357	357 tbp	0.089-0.393
	Spy	3	3 spy	0.021-0.038	19	19 spy	0.068-0.150	187	187 spy	0.099-0.280
6	**Oct**	1548	1548 oct	0.000-0.140	3317	3317 oct	0.055-0.297	3144	3144 oct	0.109-0.348
7	**Pbp**	2	2 pbp	0.063-0.064	19	19 pbp	0.080-0.168	88	88 pbp	0.109-0.351

[1] RMSD value for the *tri* geometry: 0.536 Å
[2] RSMD value for the *tri* geometry: 0.539 Å

Table 4. Results of FindGeo on the metalloprotein data set for the 35 sites in which the geometry assignment did not match that reported in (Rulisek and Vondrasek 1998). The first block includes cases where two different geometries were identified in (Rulisek and Vondrasek 1998) and by FindGeo, the second block includes cases where the geometry was classified as irregular by FindGeo but not in (Rulisek and Vondrasek 1998), and the third block includes cases where the geometry was classified as irregular in (Rulisek and Vondrasek 1998) but not by FindGeo (PDB = PDB accession code; Metal = residue name, residue number and chain identifying the metal in the PDB; Paper CG = geometry reported in (Rulisek and Vondrasek 1998); CG out = geometry identified by FindGeo; distorted geometries are indicated by "(d)"; for irregular geometries, the lowest RMSD geometry is given in parentheses; RMSD = RMSD value calculated for the geometry identified by FindGeo, in Å; for irregular geometries, the RMSD value calculated for the lowest RMSD geometry is shown; RMSD from paper CG = RMSD value calculated for the geometry reported in (Rulisek and Vondrasek 1998), in Å; distorted and irregular geometries are indicated by "(d)" and "(i)", respectively). See Supplementary Table S1 and Supplementary Figure S1 for the names of the geometries.

PDB	Metal	Paper CG	CG out	RMSD	RMSD from paper CG
1lne	CD 903 E	Octahedral	pvp (d)	0.475	0.486 (d)
1iab	CO 999 A	Trigonal bipyramidal	spy (d)	0.509	0.582 (i)
1kcw	CU 1050 A	Tetrahedral	bva	0.320	0.882 (d)
1aac	CU 107 A	Distorted tetrahedral	bva (d)	0.615	0.672 (d)
2trx	CU 109 B	Square planar	pyv	0.110	1.647 (i)
1jer	CU 138 A	Tetrahedral	bva (d)	0.551	0.591 (d)
1slv	CU 248 B	Tetrahedral (highly distorted)	bvp (d)	0.645	1.071 (i)
1occ	CU 517 A	Trigonal	tev (d)	0.659	1.039 (i)
1oac	CU 801 B	Tetrahedral	bva	0.406	0.585 (d)
1sda	CU 807 B	Square planar	bvp (d)	0.581	0.953 (d)
1aoz	CU2 702 A	Tetrahedral	bvp (d)	0.608	0.738 (d)
1aoz	CU3 702 A	Tetrahedral	bvp (d)	0.472	1.008 (i)
3pcy	HG 100 A	Tetrahedral	bva (d)	0.659	0.708 (d)
1mrr	HG 407 A	Approximately trigonal	spv	0.447	1.060 (i)
1nzr	NI 129 C	Distorted tetrahedral	bva	0.329	0.647 (d)
1frv	NI 538 B	Highly distorted tetrahedral	pyv (d)	0.292	1.133 (i)
1vkl	NI 562 A	Distorted tetrahedral	bvp (d)	0.595	0.798 (d)
1vkl	NI 562 B	Square pyramidal	tbp	0.427	0.709 (i)
1lba	ZN 151 A	Tetrahedral	bva (d)	0.443	0.774 (d)
1ali	ZN 451 A	Tetrahedral	bva	0.422	0.594 (d)
1ste	ZN 500 A	Trigonal	tev (d)	0.545	0.893 (i)
1dth	ZN 901 A	Tetrahedral	bva (d)	0.579	0.638 (d)
1ast	ZN 999 A	Tetrahedral	bva	0.394	0.680 (d)
1iag	ZN 999 A	Tetrahedral (distorted)	bva (d)	0.488	0.692 (d)
1aaz	CD 188 B	Tetrahedral (1 empty site)	irr (tev)	0.829	0.829 (i)
1lxt	CD 562 B	Distorted trigonal bipyramidal	irr (tpv)	0.675	0.768 (i)
1lne	CD 900 E	Trigonal bipyramidal	irr (tbp)	0.598	0.598 (i)
2fua	CO 216 A	Highly distorted trigonal bipyramidal	irr (spy)	0.768	0.915 (i)
1arn	CU 180	Trigonal bipyramidal	irr (tbp)	0.598	0.598 (i)

1glc	ZN 169 F	Tetrahedral (distorted)	irr (bva)	1.001	1.045 (i)
1hfc	ZN 275 A	Octahedral (1 empty site)	irr (spy)	0.640	0.640 (i)
1add	ZN 400 A	Trigonal bipyramidal	irr (tbp)	0.597	0.597 (i)
1bov	ZN 70 B	Tetrahedral (1 empty site)	irr (tev)	0.980	0.980 (i)
1occ	CU 228 B	Irregular	tet	0.410	n/a
1occ	CU 229 B	Irregular	bva	0.422	n/a

Figure 1. Two-dimensional representation of the relationships among the three possible idealized geometries (*tri*, *tev* and *spv*) for coordination number 3. Pairwise RMSD values are shown in the text box. P1, P2 and P3 indicate pathways 1, 2 and 3, respectively. See text for details.

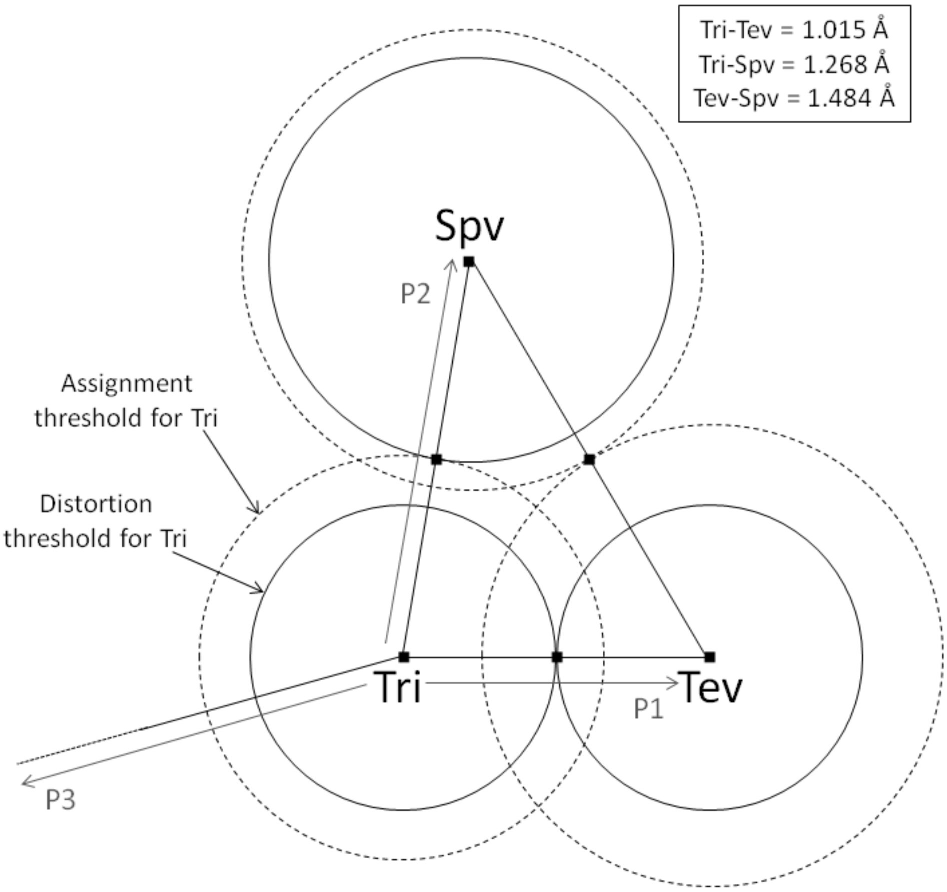


Figure 2. Distribution of the RMSD values calculated for the structures with strongly distorted *oct* geometries in the artificial data set (Panel A, 100 structures) and in the CSD-derived data set (Panel B, 3144 structures). The RMSD values calculated for individual structures are shown as grey crosses. The average RMSD value for each idealized geometry is shown as a black hyphen. Dashed grey lines correspond to the distances between the *oct* geometry and the other idealized geometries (as from Table 1).

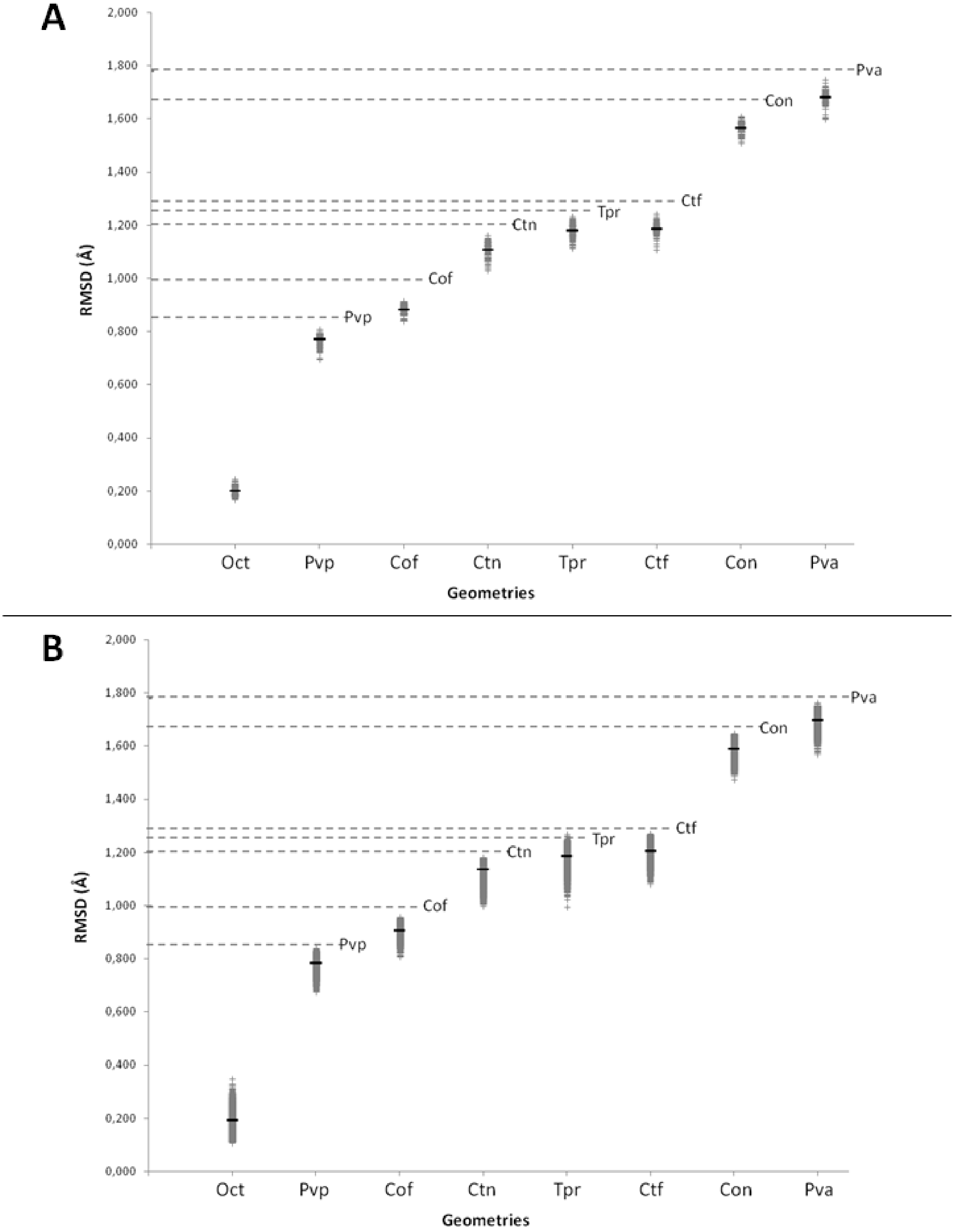


Figure 3. Geometries assigned by FindGeo in the 101 cases of the metalloprotein data set for which there was agreement with the literature (Rulisek and Vondrasek 1998) (panel A), and in the 35 cases of the metalloprotein data set for which there was not agreement with the literature (Rulisek and Vondrasek 1998) (panel B). Except for the “irregular” column, light grey sections of the columns indicate geometries classified as regular (i.e., non distorted), and dark grey sections indicate geometries classified as distorted. For the “irregular” column, the light grey, dark grey and black sections indicate geometries with coordination numbers 3, 4 and 5, respectively.

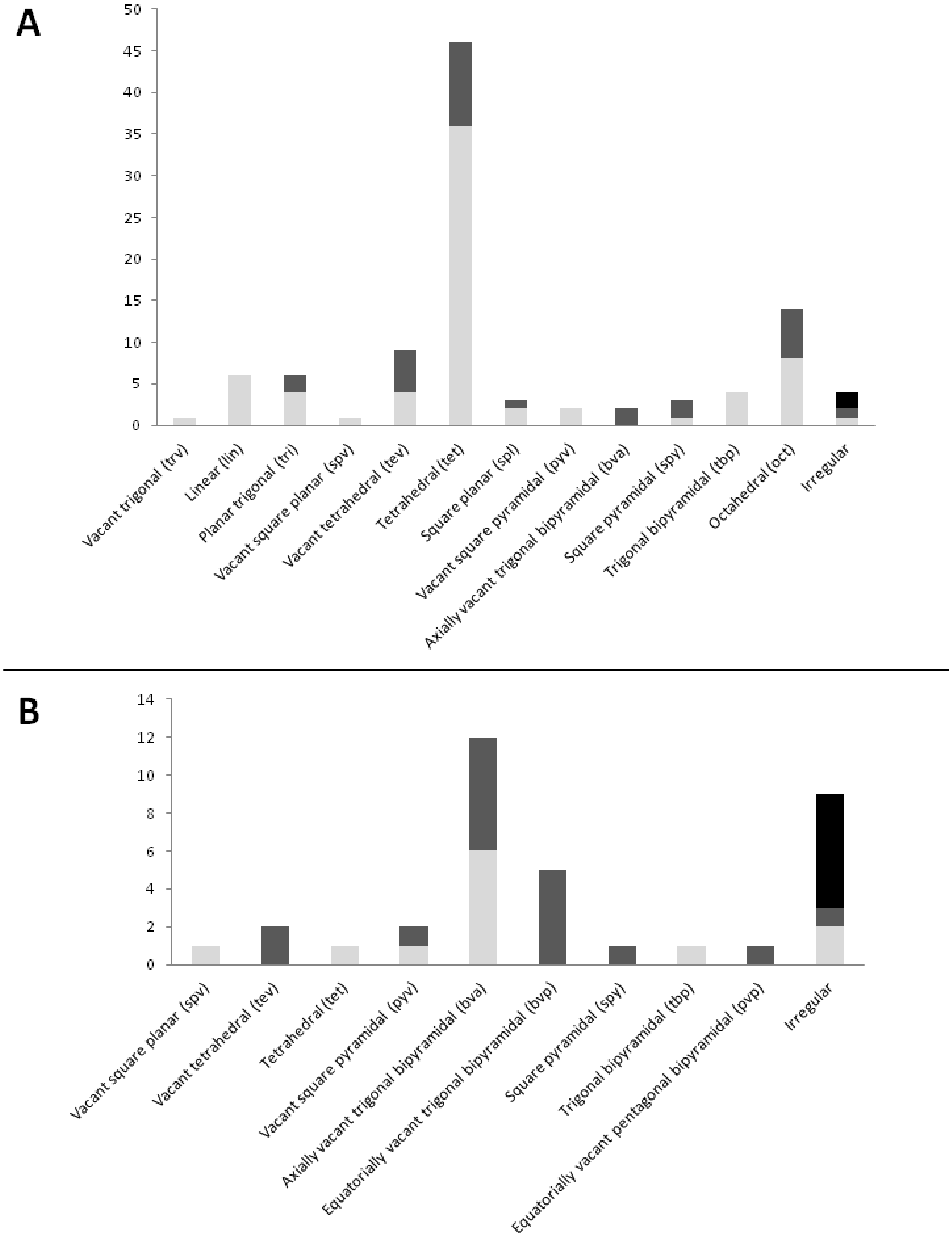
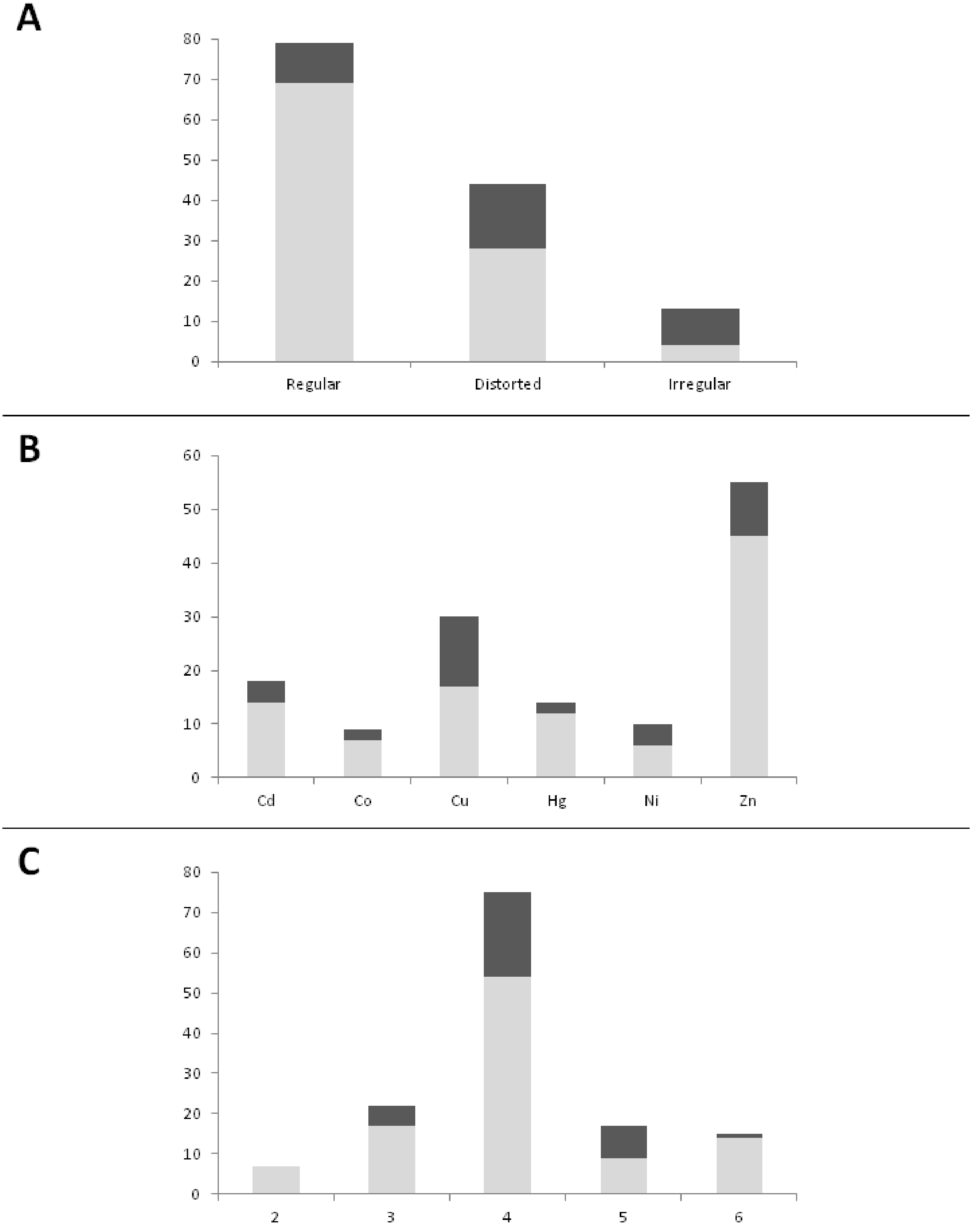


Figure 4. Number of geometries in the metalloprotein data set for which there was agreement (light grey) or not (dark grey) with the literature (Rulisek and Vondrasek 1998) in relation to whether the geometry was classified as regular (i.e., non distorted), distorted or irregular (panel A), to the metal present in the site (panel B), and to the coordination number (panel C).


A
Regular
Distorted
Irregular
B
Cd
Co
Cu
Hg
Ni
Zn
C