# Validation of metal-binding sites in macromolecular structures with the CheckMyMetal web server

Heping Zheng[1,2], Mahendra D Chordia[1,2], David R Cooper[1,2], Maksymilian Chruszcz[1–3], Peter Müller[4], George M Sheldrick[5] & Wladek Minor[1,2]

[1]Department of Molecular Physiology and Biological Physics, University of Virginia, Charlottesville, Virginia, USA. [2]Center for Structural Genomics of Infectious Diseases (CSGID) Consortium, USA. [3]Department of Chemistry and Biochemistry, University of South Carolina, Columbia, South Carolina, USA. [4]Department of Chemistry, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [5]Lehrstuhl für Strukturchemie, Universität Göttingen, Göttingen, Germany. Correspondence should be addressed to W.M. (wladek@iwonka.med.virginia.edu).

Published online 19 December 2013; doi:10.1038/nprot.2013.172

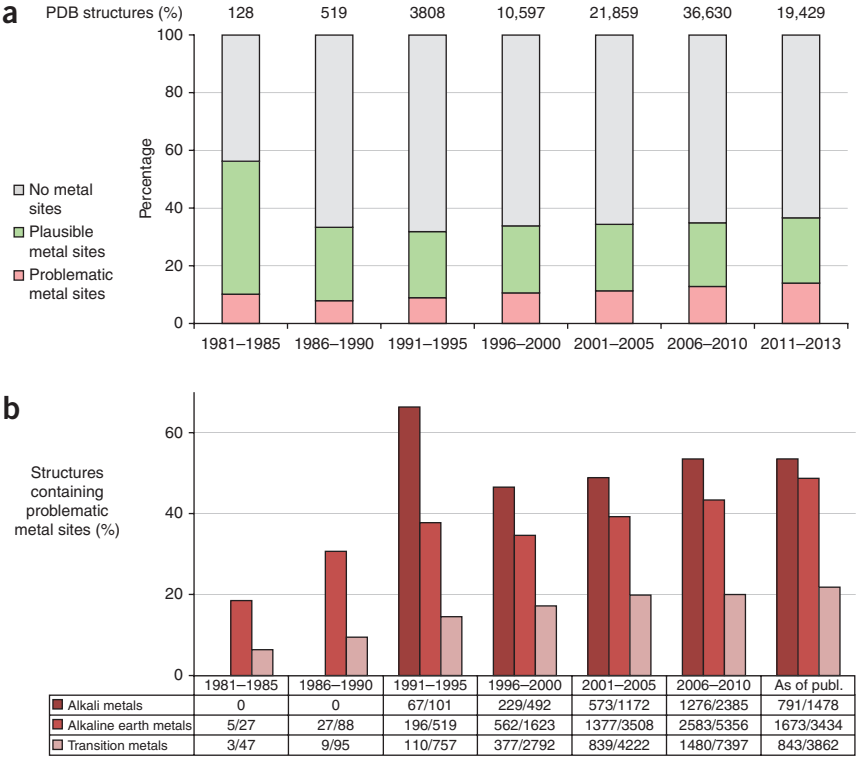
**Metals have vital roles in both the mechanism and architecture of biological macromolecules. Yet structures of metal-containing macromolecules in which metals are misidentified and/or suboptimally modeled are abundant in the Protein Data Bank (PDB). This shows the need for a diagnostic tool to identify and correct such modeling problems with metal-binding environments. The CheckMyMetal (CMM) web server (http://csgid.org/csgid/metal_sites/) is a sophisticated, user-friendly web-based method to evaluate metal-binding sites in macromolecular structures using parameters derived from 7,350 metal-binding sites observed in a benchmark data set of 2,304 high-resolution crystal structures. The protocol outlines how the CMM server can be used to detect geometric and other irregularities in the structures of metal-binding sites, as well as how it can alert researchers to potential errors in metal assignment. The protocol also gives practical guidelines for correcting problematic sites by modifying the metal-binding environment and/or redefining metal identity in the PDB file. Several examples where this has led to meaningful results are described in the ANTICIPATED RESULTS section. CMM was designed for a broad audience—biomedical researchers studying metal-containing proteins and nucleic acids—but it is equally well suited for structural biologists validating new structures during modeling or refinement. The CMM server takes the coordinates of a metal-containing macromolecule structure in the PDB format as input and responds within a few seconds for a typical protein structure with 2–5 metal sites and a few hundred amino acids.**



## INTRODUCTION

Metals are essential in many biological processes. They are present in many macromolecules, and they serve in structural and/or catalytic roles. Structural information about metal-binding environments is often used to understand the molecular mechanism of macromolecules[1]. Roughly one-third of all macromolecular structures in the PDB[2] contain one or more metal ions (**Fig. 1a**). The prevalence of misinterpretations of small molecules in the PDB was recently analyzed[3], and a substantial fraction of metal-containing structures seem to have incorrect metal assignments[4,5] owing to the ambiguity and suboptimal interpretation of the electron density map derived from macromolecular X-ray diffraction experiments[6]. Incorrect metal assignments in structures of biological macromolecules may cause the misinterpretation of molecular mechanisms and distort bioinformatics studies of living systems. Despite advances in software applications[7,8] incorporating new refinement and validation protocols[9], the modeling of metal-binding sites is not improving[1,10]. The overall trend shows that the percentage of erroneous metal-binding sites (as defined below) has stayed similar or even increased over the past 30 years (**Fig. 1a**). This phenomenon not only applies to metals commonly found in the solvent, such as sodium, but also to transition metals that are likely to be involved in catalytic reactions (**Fig. 1b**). This unexpected trend also suggests that metal-binding sites in macromolecular structures are major features that have thus far eluded the validation process. The accumulation of erroneous metal-binding sites in macromolecular structures can cause serious consequences, as it not only influences our understanding of molecular mechanisms, but also adds significant noise to data mining studies and complicates the use of prior data for knowledge-based or knowledge-assisted structure determination[11]. In addition, the scientific community needs to be aware of the abundance of errors in metal-binding sites in macromolecular structures and of the need to critically assess the quality of each metal-containing structure used.

Despite the collective experience of structural biology in determining myriad metal-containing structures, accurate identification of metals and refinement of their structural environments still pose challenges during the structure determination process[1,5]. In X-ray crystallography, metals and small molecules are usually identified by inspecting the residual electron density maps after the macromolecular model has been built[12]. Even when the unassigned electron density can only accommodate a single atom, correct interpretation of the density can still be a daunting task[13]. When observed in an electron density map, metal ions such as $Na^+$ and $Mg^{2+}$ have comparable density to that of water molecules and are often indistinguishable[14]. In NMR structure determination, experimentally derived restraints usually do not include bound metals. Modeling of metal-binding sites is then solely dependent on local geometric restraints, and therefore accurate restraint definition and validation becomes more crucial for NMR structures[15]. In addition, the X-ray electron density of noncovalently attached atoms is usually not as well-defined as the corresponding density for the macromolecule, resulting in greater difficulty in interpreting experimental data regardless of data quality[3]. As a consequence, artifacts in modeling metal-binding sites are frequent. For example, the coordination bond distances were shown to vary substantially in a resolution-dependent manner for zinc[16] and cobalt[10]. In a more elaborate model, coordination




| | 1981–1985 | 1986–1990 | 1991–1995 | 1996–2000 | 2001–2005 | 2006–2010 | As of publ. |
|---|---|---|---|---|---|---|---|
| Alkali metals | 0 | 0 | 67/101 | 229/492 | 573/1172 | 1276/2385 | 791/1478 |
| Alkaline earth metals | 5/27 | 27/88 | 196/519 | 562/1623 | 1377/3508 | 2583/5356 | 1673/3434 |
| Transition metals | 3/47 | 9/95 | 110/757 | 377/2792 | 839/4222 | 1480/7397 | 843/3862 |

**Figure 1** | Quality of metal-binding sites grouped by year for structures in the PDB, as determined by CMM. (**a**) The total number of all structures in each deposition year bin is indicated on the top of each bar. Structures without metals are shown in gray. A PDB structure is identified as problematic (red) if the number of CMM parameters (defined in text) that are outliers is higher than the number of metal-binding site(s) in that structure. Otherwise, it is flagged as plausible (green). (**b**) Percentage of structures containing problematic metal-binding sites relative to all metal-containing structures. Statistics are shown for alkali, alkaline earth and transition metals. Each cell in the table is displayed with the number of problematic metal-containing structures/the total number of metal-containing structures.

bond distances for copper[17] and cobalt[10] also appeared to be related to the specific coordination sphere composition in addition to the ligand element. This highlights the fact that correct metal placement is not trivial, and that it is highly subjective to the experience of the structural biologist building the model[6].

The most popular programs used for model refinement apply only distance restraints for each individual metal-ligand coordination bond but not global consideration of the whole metal-binding sites when run with default settings[18–21]. There is currently no consensus about how to properly define accurate restraints for metal-binding site refinement; for example, the topic frequently arises on the Collaborative Computational Project No. 4 (CCP4) Bulletin Board (http://www.ccp4.ac.uk/ccp4bb.php; a mailing list for the macromolecular crystallography community). In general, the structural community recognizes the need for structural validation, and numerous computational tools have been developed for overall validation of macromolecular structures[9,22,23]. As intended, these tools have improved the quality of structures within the PDB[9]. However, a comprehensive and easy-to-use tool specifically for validating metal-binding sites within macromolecular structures does not exist.

The identification of metals in macromolecular structures requires a systematic inspection of the entire binding environment, including the position, charge and type of atoms and residues surrounding the metal[1]. However, additional experiments are usually necessary to unambiguously verify metal identity[24,25]. Many modern synchrotrons produce X-rays at wavelengths corresponding to the absorption edges of metal ions from Mn to Zn (and beyond). Anomalous diffraction data collected at different wavelengths can be used to experimentally determine the identity of a specific ion with confidence[24]. However, insufficient anomalous scattering signal or other technical problems may preclude proper experimental identification. The CMM server is designed to systematically analyze parameters describing the metal-binding microenvironment, indicate deviations from target values and alert researchers to potential errors in metal assignment, to be used either in combination with anomalous diffraction data[24] or in situations where such data are not available.

The validation methodology used by the CMM server has been in the process of development for some time, and it has been publically available since its description in previous papers[5,26]. However, only recently was the method transformed into an easily accessible web service. Our earlier work[5,26] was mostly cited by researchers who used various aspects of the methodology to characterize the identity of metal ions in macromolecular structures[27–29]. The fully developed CMM server is routinely used to validate newly determined structures by the CSGID and the Midwest Center for Structural Genomics[30], and it has been cited by other laboratories[31].

## The CMM server

The CMM web server is freely available at http://www.csgid.org/csgid/metal_sites. CMM is able to identify and analyze all metal-binding sites when it is provided with a PDB identifier or coordinates in the PDB file format (**Fig. 2**). The output is shown as a table, with one metal-binding site per row (**Fig. 3**). CMM uses up to eight parameters to evaluate the inherent consistency and geometrical arrangement of each metal-binding site. Six parameters are dependent on the coordination sphere geometry and valence[32] of each binding site, as introduced in **Box 1**. Two additional X-ray crystallography–specific parameters are related to atomic displacement factors (temperature or *B-factors*) and *occupancy* (parameter names are in plain *italics*). The threshold values for all parameters are determined by using a benchmark data set, as described in the Experimental Design section. The values of the six (or eight) parameters are shown and each is labeled as 'outlier,' 'borderline' or 'acceptable' by coloring each table cell as red, amber or green (RAG), respectively. For those who may have difficulty discerning the RAG colors, 'outlier' parameters are also shown in **underlined bold**, and 'borderline' parameters are

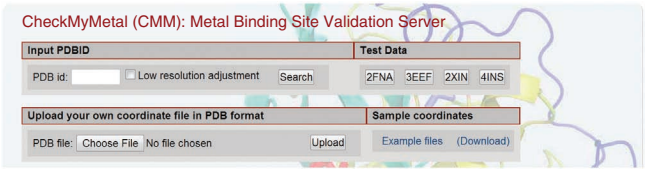



**Figure 2** | Job submission interface of the CheckMyMetal (CMM) server.

*underlined italics*. Table cells that are not applicable are colored gray. The RAG color coding provides a quick overview of the quality of each metal-binding site.

When CMM detects significant deviations of one or more parameters from target values, it may suggest a ranked list of probable alternative metals that are more consistent with the structural environment of the analyzed ion. Alternative metal ion(s), if applicable, are listed for each problematic metal-binding site. The results page includes a Jmol applet (which requires the Java plug-in) to view metal sites in 3D[33] (**Fig. 3b**). Symmetry-related ligands are marked with the prefix 'sym-' in the Jmol applet, and they can be switched on/off by using a radio button. The first column of the output table contains buttons to switch which metal site within a structure is displayed in the Jmol window.

The elemental composition of the metal-binding coordination sphere (i.e., all nonhydrogen atoms that serve as metal ligands grouped by atomic element) is shown together with graphs plotting metal-ligand distance distributions for each individual ligand element (**Fig. 3e**). These graphs show both the distribution of metal-ligand distances in the current structure and the distribution of the distances observed in the Cambridge Structural Database (CSD), a database of very high–resolution X-ray crystallography structures of small molecules[34].

### Applications of the CMM server

CMM is generally applicable to any metal-containing macromolecular structure, including complexes with nucleic acids and/or polysaccharide chains. As long as the structure is in PDB format, it may be evaluated by the CMM server regardless of the method used to determine the structure (this includes theoretical models). All metal-containing macromolecular structures in the PDB released as of May 2011 were evaluated by CMM. Numerous major errors in either assignment of metal and/or environment were detected, as summarized in **Figure 1**. Several examples are shown in the ANTICIPATED RESULTS section to demonstrate CMM's capability to handle metal binding sites in structures determined by X-ray crystallography and NMR, which comprise >99% of PDB structures.

CMM is applicable not only to structural biology but also to many other research areas, including biochemistry, molecular biology, computational biology and drug discovery. With CMM, structural biologists can identify and correct metal-binding site errors in their own structures while still refining the structure and before depositing it into the PDB. However, most researchers from other fields use structures already in the PDB as the starting point for their research. The errors in metal-binding sites in the PDB make validation prudent before investing time and resources on subsequent experiments. To this end, CMM provides a do-it-yourself method for evaluating the quality of the metal-binding sites in macromolecular structures of interest. Last but not least, the principles of CMM may be applied

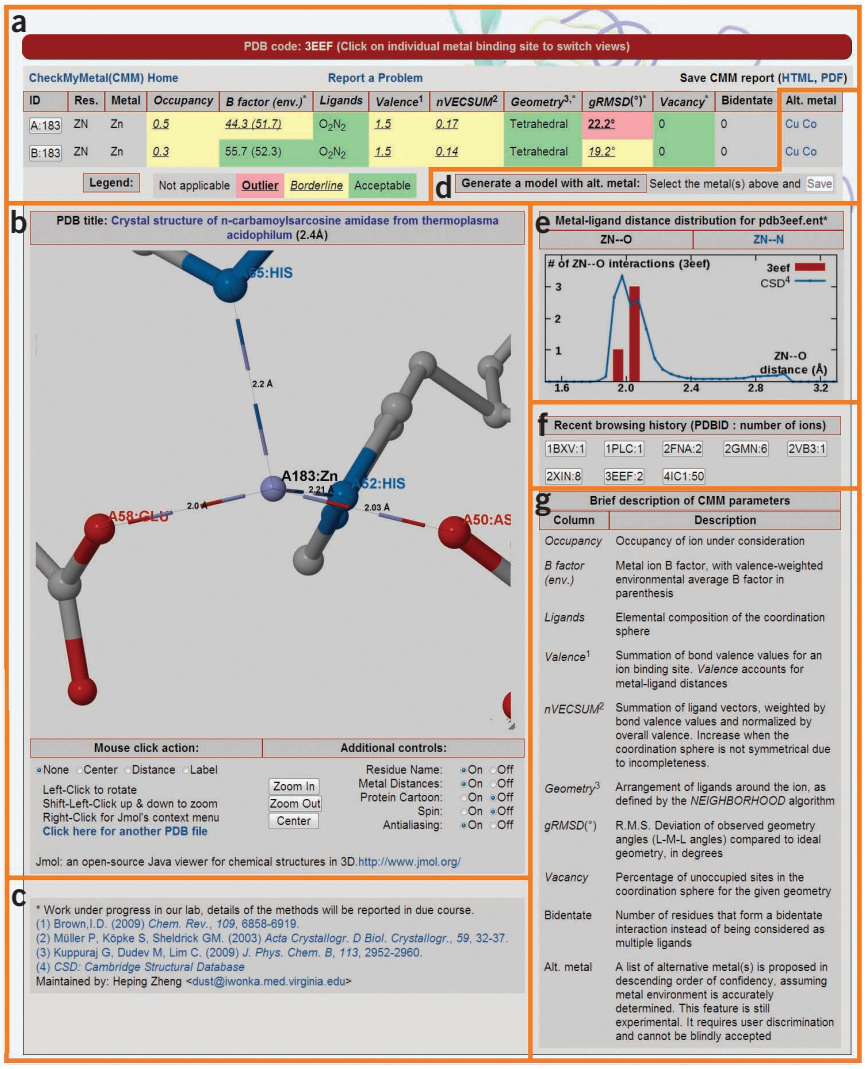



**Figure 3** | CheckMyMetal (CMM) server: screenshot of the various components of the results display interface. (**a**) Summary table with one metal-binding site per row. Residue ID, type of metal and all parameters evaluated are shown for each metal-binding site individually. (**b**) Interactive metal-binding site viewer implemented using Jmol. (**c**) References. (**d**) Utility to retrieve a model containing alternative metal ion(s). (**e**) Metal-ligand distribution of the analyzed structure (shown as red columns) and of that from CSD (shown as blue lines). (**f**) Shortcuts to retrieve recent jobs submitted from the same client computer. (**g**) Brief descriptions of the columns listed in the results summary table.

## Box 1 | The six derived parameters: *ligands*, *valence*, *nVECSUM*, *geometry*, *gRMSD* and *vacancy*

1. The *ligands* parameter describes the elemental composition[7,11] of the first coordination sphere.
2. The overall bond *valence* parameter is the summation of individual bond valence values[10] for all metal-ligand bonds.
3. The *nVECSUM* parameter is a valence-normalized adaptation of VECSUM, which is a vector-based summation of bond valence vectors[12], and it is valid for metal sites that display symmetrical coordination geometries (see Experimental design).

**! CAUTION** The *valence* and *nVECSUM* parameters should be interpreted with care owing to the typical resolutions of macromolecular structures and other possible complications. *nVECSUM* is not applicable to asymmetrical geometries, such as those with stereochemically active lone pairs (see Experimental design).

4. The *geometry*[11,13] parameter describes the pattern of 3D arrangement of ligands around the metal as used in coordination chemistry.
5. The *gRMSD* parameter measures the overall deviation of all LML angles from idealized coordination geometry. (Deviations from ideal metal-ligand bond distances are measured by the overall *valence* and *nVECSUM* parameters.)
6. The *vacancy* parameter measures the percentage of vacant coordination sites for a specific *geometry*.

to create more accurate algorithms for the modeling of metal-binding sites for structure refinement, computational biology and *ab initio* structure modeling.

### Comparison with other methods

A few other programs or services for analyzing the metal-binding sites in metalloproteins have been described in the literature, each with a different emphasis. These programs and services, including CMM, are listed in **Supplementary Table 1**, along with a comparison of the features of each. Programs that predict metal-binding sites include FINDSITE-metal[35], MetSite[36], SVM-Prot[37], SeqCHED[38] and metalDetector[39]. Databases for querying metal-binding sites include MESPEUS[16], MIPS[40], MDB[41], MetalPDB[42], Metal-MACiE[43], MINAS (http://www.minas.uzh.ch/) and PROMISE[44]. The services PDBsum[45] and PDBeMotif[46] include additional annotations. The visualization software UCSF Chimera[47] implements a structure analysis tool for metal geometry. In general, CMM complements these existing programs and servers; for example, it does not predict new metal-binding sites or provide search functionality. To our knowledge, CMM is the first service to implement a comprehensive validation mechanism by identifying and flagging problematic metal-binding sites to verify that metal-binding sites are modeled as accurately as possible. Moreover, as part of an overall research project to investigate metal-containing proteins, CMM might be used in conjunction with other programs that can predict metal-binding sites such as FINDSITE-metal[35] or programs that provide search functionality such as MetalPDB[42].

### Experimental design

**CMM server backend.** The CMM server backend uses an enhanced version of the previously described NEIGHBORHOOD SQL database[5], which stores PDB-derived information of all modeled metal ions, their neighboring atoms and residues, together with each coordination bond as a vector. The relational database provides an effective way to query and classify a very large set of metal ions and their coordinating ligands for further analysis of the metal-binding sites in a specific structure. The CMM-related parameters are derived and stored in the database for each metal-binding site, whereas the web interface is rendered using CakePHP after querying the database. Structures that are uploaded to the server are processed on the fly, and they are discarded 24 h after analysis. The results are only available to the submitter through a passcode that is delivered to the identified IP address submitting the original job. The bond valence values, geometry data and other characteristics of the coordination sphere (e.g., bidentate interactions, coordinating atoms and residues) are calculated for each metal-binding site by CMM.

***Valence* and *nVECSUM*.** The overall bond *valence* refers to the sum of the individual bond valence values $v_i$ for each metal-ligand interaction within the coordination sphere according to the bond valence model[32]. For an ideal site, the value of the valence parameter should be equal to the oxidation number of the metal—e.g., a $Zn^{2+}$ site ideally would have a valence of 2.0. Each bond valence is calculated from each observed metal-ligand distance $R_i$ by $v_i = \exp((R_o - R_i)/b)$, where $R_o$ is a constant (for a given metal element, oxidation state and ligand element) describing the ideal distance where the bond valence is 1, and $b$ is an empirical constant. The bond valence $R_0$ values reported previously were used for valence calculations[48,49]. The bond valence $R_0$ values were reported differently in these publications[48,49], and the use of the most accurate $R_0$ values is crucial. For example, the calculated overall bond valence that reflects copper oxidation state differs significantly if the updated $R_0$ values[49] are used in the place of the originally reported $R_0$ values[48]. The *nVECSUM* parameter is defined as the amplitude of the vector sum of bond valence vectors normalized by the overall *valence*. The bond valence vector for each atom within 4 Å of the metal is defined as a vector with magnitude equal to the bond valence and directed along the metal-ligand bond (originating at the metal). For an ideal site with perfect symmetry, identical ligands and bond lengths, all of the individual bond valence vectors would cancel each other out and *nVECSUM* would be 0. *nVECSUM* is a valence-normalized adaptation of the VECSUM algorithm[26]. Thus, both the overall *valence* and *nVECSUM* computations provide two single parameters summarizing deviations in symmetry and bond distance from ideal geometry for all of the metal-ligand interactions[26]. The validity of the *valence* and *nVECSUM* parameters is further verified by the agreement between the valence distribution peaks and the oxidation states (**Supplementary Fig. 1**) and *nVECSUM* values close to 0 (**Supplementary Fig. 2**) for high-resolution metal-containing X-ray structures from the PDB.

VECSUM analysis assumes that ligands are symmetrically arranged in the first coordination sphere; therefore, it is not applicable for sites with asymmetrical ligand arrangement owing to the presence of stereochemically active lone pairs (e.g., Pb(II) or Sn(II)), or for sites where not all of the expected inner-sphere ligands are modeled. Although the *nVECSUM* calculation in CMM checks for missing atoms in the coordination shell for the most common metal sites that display symmetrical coordination geometries, the validity of this method may be compromised to some extent because of imprecise interatomic distances between a metal and its ligands for macromolecular structures determined at medium to low resolution. Additional complications such as a mixture of metal ion composition may also be introduced during nonphysiological sample manipulations, including expression in non-native organisms, biochemical experiments and radiation damage. In these situations, it is possible that metal sites may be occupied by a mixture of ions owing to variations in the supply of the optimal ion or ions, being partially replaced by presumably 'unnatural' alternative ions. For metal ions that can have more than one oxidation state, radiation damage can very rapidly reduce metal ions (e.g., $Fe^{3+}$ to $Fe^{2+}$, $Cu^{2+}$ to $Cu^{+}$), resulting in slight expansion or contraction of the coordination sphere to balance the overall valence even at cryogenic temperatures[50,51]. The bond valence model has the limitation that it is restricted to localized bonds and does not apply to situations involving metal-metal bonding or appreciable π-backbonding[32]. Yet cation-π interactions that coordinate the d-orbital of transition metals that are well known in small organic compounds are rarely observed for aromatic rings from protein side chains (Phe, Tyr and Trp). Thus, both the *valence* and *nVECSUM* parameters should be carefully interpreted when another metal ion or another unusual ligand with delocalized electrons is observed in the metal surrounding environment.

**Geometry assignment.** The cutoff distance for the first coordination sphere is determined by the individual bond valence value that is proportional to the ratio of the assumed oxidation state (as implied by the residue name in the PDB file) to the coordination number (i.e., the number of ligands coordinating a given metal in the first coordination sphere). Currently, CMM includes only the most commonly observed atoms (N, O and S) in its calculations. For all atoms identified as part of the coordination sphere, CMM uses a novel algorithm to compare the geometry of the observed ligand-metal-ligand (LML) angles to the idealized LML angles for reference geometries. Thirteen reference geometries are considered for sites with observed coordination numbers between two and eight (linear, triangle, square planar, tetrahedral, triangle bipyramid, octahedral, trigonal biprism, trigonal antiprism, capped trigonal biprism, pentagonal bipyramid, cubic, square antiprism and dodecahedral). The *gRMSD* parameter is defined as the r.m.s.d. of observed LML angles as compared with their idealized values. For each metal-binding site, the geometry deviation value is calculated for each possible reference geometry. A weighted function that incorporates both the geometric deviation and penalties for vacant coordination sites and uncommon geometries is used to determine the best-matched geometry.

**Bidentate coordination geometry.** It has been shown previously that calcium usually adopts an octahedral geometry with coordination number 6, and that the majority of calcium-binding sites with coordination number 7 or 8 are attributable to bidentate interactions with Asp or Glu, which could be considered as occupying one coordination site[5,52]. This phenomenon may apply to heavier alkaline earth metals ($Sr^{2+}$ and $Ba^{2+}$), but their rare occurrence in the PDB precludes a conclusive proposition. In the current study, all potential bidentate interactions are evaluated to determine whether they should be treated as a bidentate coordination or as two separate coordinations. The number of vacant coordination sites is deduced once the best-fitting geometry is chosen. The *vacancy* parameter is then calculated as the percentage of these vacant sites divided by the coordination number for the chosen *geometry*. If there is a potential bidentate interaction involving two atoms (e.g., carboxylic acid, guanidine) and the ion, the geometry search algorithm calculates the *gRMSD* for both cases, by treating the potential bidentate coordination either as one 'pseudo-atom' or two independent coordinating atoms. CMM then selects the case with the lower *gRMSD*. When multiple potential bidentate ligands are present around the metal-binding site, all combinations of bidentate and nonbidentate interactions are considered and the combination that gives the lowest overall *gRMSD* is selected.

**Classification of parameter values.** A benchmark data set consisting of high-resolution (≤1.5 Å), metal-containing X-ray structures from the PDB[2] were used to determine the thresholds for classification of the six derived parameters. A few structures (3fiy, 3i24, 3l52 and 3keo) were excluded owing to unusually (and suspiciously) large numbers of uncoordinated metal ions[4]. This yields a data set of 7,350 metal-binding sites in 2,304 macromolecular structures. The threshold values were determined by using metal-binding sites in protein structures as a benchmark data set rather than, say, higher-resolution data from small-molecule structures. This permitted the inclusion of metal-binding sites that deviate from ideality owing to strains associated with the protein backbone. For each parameter, three different zones are assigned as outlier (red), borderline (amber) and acceptable (green) to indicate the deviation from that parameter's target values. The thresholds for classifying coordination sphere composition and geometries into either outlier, borderline or acceptable zones are based on statistics described previously[5,53,54] (http://tanna.bch.ed.ac.uk/qg3.htm). The thresholds for overall valence, nVECSUM, gRMSD and vacancy parameters were empirically selected on the basis of the distributions of these parameters in the benchmark data set (**Supplementary Figs. 1** and **2**). All distributions featured a shape with the steepest edge of the major peak at around half the height of each peak, and background noise at around 10% of the peak height. Thus, for each of the four parameters, the regions where the height of the distributions were either >50%, between 10 and 50% or <10% of the height of the distribution peak were defined as the acceptable, borderline and outlier zones, respectively. Thus, the approximate values of the parameters that bracket these zones on the distributions are used as the threshold values for classification (**Supplementary Table 2**).

For the overall valence parameter, multiple distributions were used, subdivided both by metal identity and assumed oxidation state (valence): +1 ($Na^{+}$/$K^{+}$/$Cu^{+}$), +2 ($Mg^{2+}$/$Ca^{2+}$/$Mn^{2+}$/$Fe^{2+}$/$Co^{2+}$/$Ni^{2+}$/$Cu^{2+}$/$Zn^{2+}$) and +3 ($Fe^{3+}$/$Co^{3+}$/$Ni^{3+}$). Although we cannot exclude the possibility of the presence of $Ni^{3+}$, the +3 oxidation

state of nickel (shown as a minor peak in **Supplementary Fig. 1**) is only rarely observed in macromolecular structure, and we encourage further experimental confirmation of the oxidation state before assigning a nickel ion as $Ni^{3+}$. For each metal, borderline and outlier zones were defined symmetrically both above and below the acceptable range for the corresponding valence. For all distributions that involve two peaks, thresholds for both peaks were used to assign acceptable, borderline and outlier zones bimodally. In some cases, the distributions for two different metals with the same oxidation state are similar enough for the same thresholds to be used (e.g., $Ca^{2+}$ and $Mg^{2+}$; **Supplementary Fig. 2** and **Supplementary Table 2**). Some peaks were too poorly resolved to infer reliable threshold values, particularly the peaks representing less common oxidation states of metals able to adopt multiple states (e.g., $Fe^{2+}/Cu^{2+}/Ni^{3+}/Co^{3+}$), and thus the same threshold values determined for ions with the same overall valence are used (**Supplementary Table 2**).

For the *nVECSUM*, *gRMSD* and *vacancy* parameters, the distributions all monotonically decrease, except at values close to 0 (and thus ideality). Therefore, a single threshold can be used to define the borderline and outlier zones. The borderline and outlier thresholds are >0.10 and >0.23, respectively, for *nVECSUM*, >13.5° and >21.5°, respectively, for *gRMSD* and >10% and >25%, respectively, for the *vacancy* parameter, which is the percentage of all expected coordination sites left vacant (**Supplementary Fig. 2** and **Supplementary Table 2**). For example, ions with all coordination sites occupied by ligands (*vacancy* = 0) are classified as acceptable. For geometry with an expected coordination number greater than four, metals with one vacant coordination site (*vacancy* ≤ 25%) are borderline, and metals with two or more vacant coordination sites (*vacancy* > 25%) are outliers (**Supplementary Table 2**). Although the same thresholds are used regardless of resolution, low-resolution structures often contain fewer ordered water molecules, and hence one may observe fewer coordinating ligands[5].

**X-ray crystallography–specific parameters—*occupancy* and *B-factor*.** Metal occupancy is usually expected to be 1.0. Thus, only full *occupancy* is defined as acceptable, partial occupancy is borderline and essentially zero occupancy (<0.01) is outlier. The B-factor parameter shows two values: the B-factor of the metal atom and the 'environmental' B-factor, which is the bond-valence–weighted mean of the B-factors of all ligand atoms. When the occupancy is >0, the ratio between the metal B-factor ($B_{met}$) and the environmental B-factor of the coordinating atoms ($B_{env}$) is defined as $\exp(-|\ln(B_{met}/B_{env})|)$. By this definition, the ratio is 1 when $B_{met} = B_{env}$, and decreases as the absolute difference between $B_{met}$ and $B_{env}$ increases. The distribution of *B-factor* ratios for all metal-binding sites in the benchmark set were used to generate threshold values for borderline (<0.86) or outlier (<0.54) *B-factor* ratios in a manner similar to that described in the previous paragraph (i.e., by finding the points in the distribution where the distribution was 50 or 10% of the peak height; **Supplementary Fig. 2**).

**Ligands generated by symmetry operations.** For structures determined by X-ray crystallography, some atoms in the first coordination sphere of a metal may not be found in the asymmetric unit, but instead are provided by one or more symmetry-related copies of the model. When application of crystallographic symmetry operators may be needed to complete the coordination sphere for a metal-binding site, symmetry-related ligands are detected by the CONTACT program from the CCP4 suite[55].

**Alternative metal suggestion.** CMM will suggest a list of alternative metals when it suspects an incorrect metal assignment. The potential replacement metal ions are ranked using a scoring function, which measures the best fit of each to the metal ion. The scoring function is empirically derived from calcium bond valence sum (CBVS) analysis[26]. The CBVS value is calculated as the overall valence when a 'pseudo-calcium ion' replaces the assigned metal. The resulting CBVS is then compared with characteristic values for different metal types to calculate the CBVS deviation for each metal. Metals with reasonable CBVS deviations are reported by the server as a potential metal that can accommodate the observed coordination geometry. If no alternative metals are suggested and multiple parameters are marked as outliers, one is urged to re-examine the metal-binding environment carefully. The current method distinguishes between alkali and alkaline earth metals well, but it may have difficulty in correctly identifying a proper transition metal. The alternative metal suggestion algorithm is also limited by the existing arrangement of ligands, which may itself be wrong. In such situations, the metal-binding environment needs to be further refined before CMM can suggest potential alternative metals. For the best results, each metal in the alternative metal list should be modeled in the site, and structures should be re-refined and resubmitted to CMM.

## MATERIALS

### EQUIPMENT

**Data**

- The coordinates of the metal-containing macromolecule(s) or the coordinates of the metal and all of its ligands in PDB format: for X-ray crystallography structures, the PDB file must include a 'CRYST1' record indicating the unit cell dimensions and crystallographic space group, as specified by the PDB format (http://www.wwpdb.org/docs.html) in order to ensure the identification of all coordinating ligands for metal ions located near the border of the crystallographic asymmetric unit
- (Alternative) A PDB identifier can be used instead of a PDB file if the metal-containing structure is already in the PDB

**Hardware and software**

- A computer with Internet connectivity
- A two-button mouse. A three-button mouse, especially one with a middle button scroll wheel, is recommended for visualization using Jmol[33]
- A relatively modern web browser (supporting DOM Level 2, CSS2 and XHTML 1.0) with JavaScript enabled. CMM has been tested to be compatible with Mozilla Firefox (version 3.0+), Google Chrome (any version), Microsoft Internet Explorer (version 6.0+), Apple Safari (version 4.0+) and Opera (version 8.0+)
- A Java plug-in for the web browser, for interactive viewing of the metal-binding site using Jmol[33]
- (Optional) Text editor (e.g., Wordpad, TextEdit, Vim, Emacs) for manual editing of the PDB format file, needed only when the identity of the metal is problematic
- (Optional) Interactive modeling software (e.g., COOT[56]) and a crystallographic refinement program (e.g., REFMAC[18]), for further remodeling of metal binding is desired in cases where an error was found in addition to the evaluation of the metal-binding site

## PROCEDURE

### Submitting a job ● TIMING ~10 min

**1|** Open the CMM home page at http://www.csgid.org/csgid/metal_sites in a web browser.

**2|** *Prepare or select the data*. This step differs depending on whether an existing structure in the PDB (option A) or a user-generated PDB file (option B) is to be analyzed (**Fig. 2**). Some example structures are provided as buttons labeled with the identifiers of selected PDB structures under the 'test data' section, which demonstrate the use of the server with a single click.

(**A**) **Specify a structure already deposited in the PDB by ID**

(i) Input the four-character PDB code in the 'PDB id:' field (**Fig. 2**).

(ii) (Optional) Check the 'Low resolution adjustment' checkbox to instruct the server to use more generous thresholds, which may be more appropriate for structures determined by X-ray crystallography from low-to-medium–resolution data. Using the 'Low resolution adjustment' option will not change the calculated parameters for each site, but it will change the thresholds used to distinguish the outlier-borderline-acceptable ranges as indicated by RAG background colors. This option will only affect structures with a resolution worse than 2 Å (refs. 10,16).

▲ **CRITICAL STEP** The authors suggest that less-experienced users not use this option. As metal coordination chemistry itself obviously does not depend on the resolution of structural data, the consideration of resolution in the evaluation of metal-binding site quality is questionable in itself. However, having a more generous threshold could be informative for expert users to identify relatively good metal-binding sites in a low-resolution structure. Therefore, this feature is considered for informational purposes only, and the resulting validation report should not be compared with other validations. When this option is chosen, the results should be interpreted with great care.

(iii) Press the 'Search' button.

**? TROUBLESHOOTING**

(**B**) **Upload a coordinate file in PDB format**

(i) Press the button labeled either 'Choose File' or 'Browse' (this may vary depending on the web browser; **Fig. 2**) to open a file selection dialog.

(ii) Select a coordinate file with the *.pdb extension. The name of the file selected will be displayed, possibly abbreviated.

(iii) Press the 'Upload' button.

**? TROUBLESHOOTING**

### Monitoring the job ● TIMING ~1 min

**3|** Wait for the server to respond with the validation results, which normally will happen within a few seconds.

**? TROUBLESHOOTING**

**4|** (Optional) Press one of the buttons in the section 'Review recent browsing history' below the Jmol window as a shortcut to retrieve a recently submitted job (**Fig. 3f**). The text on the button is in the format of 'XXXX:Y', where 'XXXX' is the PDB code (or a temporary passcode in the case of an uploaded structure) and 'Y' is the number of metal-binding sites detected in the structure. Jobs are stored temporarily on the server and available for the submitter to retrieve within 24 h by using the PDB code or the assigned passcode.

### Viewing validation results ● TIMING ~5 min

**5|** The results table will be shown, with one metal-binding site per row (**Fig. 3a**). The table contains 13 columns. The first column of the output table is a button for selecting a particular site (see Step 7), followed by the residue name and atom name of the metal in the second and third column, respectively. The remaining ten columns define the properties of the metal-binding site, including the six derived parameters (**Box 1**), two additional X-ray crystallography–specific parameters (if appropriate), the number of bidentate interactions and a list of alternative metal(s), if any. A brief description of the meaning of each column is displayed on the bottom right (**Fig. 3g**), along with references (**Fig. 3c**). A 'save CMM report' feature is provided at the top right corner of the results table, which allows saving the validation report in either HTML or PDF format. For a modeled metal-binding site with no errors, all table cells in the given row will be shaded green (or gray if certain parameters are not applicable). If CMM detects no modeling errors in all metal-binding sites, Step 10 and onward do not apply, whereas Steps 6–9 are optional. Otherwise, each metal-binding site should be analyzed as described below.

**? TROUBLESHOOTING**

**6|** Examine the distance distribution for different types of metal-ligand interactions (**Fig. 3e**). If two or more types of metal-ligand interactions exist in the structure being analyzed, click on the label of each type of metal-ligand interaction to toggle among the different distribution graphs. The label of the current distribution graph will be shown in black. On each graph, the normalized distribution of metal-ligand interaction distances found in CSD is shown by blue lines,

and the distribution of distances in the current structure is shown by red boxes. The horizontal axis enumerates the metal-ligand distance and the vertical axis enumerates the number of interactions in each bin for the current structure. These two distributions should be similar for structures with reasonable metal-ligand distances, which is one indicator of properly modeled metal-binding sites. Outliers indicate the potential presence of problematic metal-binding sites that may need closer inspection.
**? TROUBLESHOOTING**

**7|** Press the button corresponding to the metal site of interest in the first column of the results summary table (**Fig. 3a**). The text in the button is in the format of 'X:YYYY', where 'X' is the chain letter and 'YYYY' the numeric residue ID of the metal ion. The Jmol window (**Fig. 3b**) will recenter on the metal ion and binding site of interest.
**? TROUBLESHOOTING**

**8|** Inspect the metal-binding site in the Jmol window to identify potential problems (**Fig. 3b**). One may use the mouse to navigate around the metal site in 3D: left-click and drag to rotate about the metal ion, hold shift and left-click and drag up or down to zoom, and right-click to bring up the Jmol context menu. These instructions are summarized in the 'Basic controls' box below the Jmol window. The size of the Jmol window can be adjusted by using the gray triangle near the Jmol logo at the bottom right corner. Detailed instructions about the use of Jmol may be found on the Jmol website (http://jmol.sourceforge.net/).
**? TROUBLESHOOTING**

**9|** By default, only the metal ion and the residues coordinating it are shown (in stick representation), but additional controls below the Jmol window may be used to customize the display of the metal-binding site (**Fig. 3b**). 'Residue name' and 'metal distances' can be toggled off for a clearer view of a complex site. 'Protein cartoon' renders the protein secondary structure in cartoon representation, and 'Spin' slowly spins the structure about a vertical axis. 'Antialiasing' controls whether the 3D structure is displayed with smoothed edges; toggling it off will lower the quality of the display but may enhance performance.
**? TROUBLESHOOTING**

**Analyzing problematic sites ● TIMING ~10 min**
**▲ CRITICAL** All of the following analyses (Steps 10–19) should be performed for one metal-binding site at a time. Each problematic metal-binding site needs to be further investigated individually.
**▲ CRITICAL** Detailed definitions and complications of all parameters are given in the 'Experimental design' section. Brief descriptions of each parameter are also given in **Box 1**. Read those definitions carefully before proceeding with analysis of problematic sites (Steps 10–15).

**10|** (Optional) Analyze problems in *occupancy* and *B-factor* agreement. Metal-binding sites with partial occupancy do occur, but they should always be carefully considered. Large deviations between the metal B-factor and environmental B-factors usually reveal the incorrect metal identity.
**▲ CRITICAL STEP** The crystallographic B-factor and occupancy of an atom are correlated. If the metal identity is certain (e.g., by spectroscopic or biochemical experiments), but large deviations between metal B-factor and environmental B-factor are still observed, the accuracy of the occupancy should be considered for further analysis.

**11|** Analyze problems of unusual geometry. The presence of uncommon geometry is rather rare; therefore, the geometry should usually be correct for sites with three or more ligands. However, it is very common to have 'poorly coordinated' sites where only one ligand is identified in close vicinity. In these cases, the *geometry* parameter cannot be determined at all. The *geometry* type is only determined for sites with a minimum of two ligands. Sites with only two ligands are generally unreliable (except for linear geometry), and they are indicative of an incompletely modeled metal-binding environment.

**12|** Analyze problems of geometric irregularity with *gRMSD*. The *gRMSD* parameter should be evaluated only when the geometry type is correct. Owing to backbone restrictions in proteins, it is not uncommon to observe some distortions in the angles from ideality (**Supplementary Table 2**). *gRMSD* deviations of >20° usually indicate problems in the metal-binding environment.
**! CAUTION** Geometries with one or two ligands are usually significantly incomplete, and the *gRMSD* will not be meaningful: *gRMSD* cannot be calculated at all for sites with a single ligand, and it only incorporates a single angle in sites with two ligands. As *gRMSD* is a parameter dependent on and directly derived from the *geometry* parameter, the calculation of the *gRMSD* value differs completely for different *geometry* types. As a guideline, the common *geometry* type is correct if the *geometry* parameter is in either the acceptable or borderline range, as described in Step 11.

**13|** Analyze problems involving an incomplete coordinate sphere using *vacancy* and *nVECSUM*. Similarly to *gRMSD*, the *vacancy* parameter is only meaningful when the *geometry* is correctly identified. For sites with three or more ligands, the *vacancy* parameter explicitly measures the percentage of expected ligands missing. A nonzero vacancy value, by definition, indicates an incompletely modeled metal-binding environment. The *nVECSUM* parameter can be used to evaluate the symmetry of the ligand arrangement around the metal. Large *nVECSUM* values reveal asymmetrically modeled ligands around the metal, which usually indicates missing ligand(s) in the first coordination sphere.
▲ **CRITICAL STEP** The *geometry* parameter can effectively evaluate the completeness of the coordination sphere for sites with one or two ligands. The completeness of sites with three or more ligands should be determined more accurately by using the *vacancy* parameter. The *nVECSUM* parameter can be used in all cases to indicate completeness, in conjunction with either the geometry parameter or the *vacancy* parameter.

**14|** *Analyze valence problems*. For a site with complete coordination sphere, large deviations in the overall valence parameter usually indicate an incorrect metal identity. Consult the metal-ligand distance distribution chart(s) described in Step 6 to further spot potential problems (**Fig. 3e**). A single outlier in the distribution of metal-ligand distance usually indicates an error in the metal-binding environment, whereas multiple outliers are more likely to indicate an error in metal identity assignment.
**!** **CAUTION** The use of the *valence* and *nVECSUM* parameters may not be applicable in specialized cases when metal-metal bonding or appreciable π-backbonding is present in the metal-binding environment. In such cases, the reported *valence* and *nVECSUM* parameters should be interpreted with great care, and outlier values do not necessarily indicate defects in the metal-binding site modeling.
▲ **CRITICAL STEP** The overall valence parameter evaluates the combined effect of both the metal-ligand distances and the completeness of the coordination sphere. This parameter is accurate only for sites with a complete coordination sphere, and it will not be meaningful for sites with missing ligands.

**15|** Analyze problems in coordination sphere composition. When there is a problem in ligand identity, check the orientations of His, Asn and Gln side chains and the chirality of small molecules adjacent to the metal ion (if appropriate). Unfavorable metal-ligand interaction(s) are usually related to problems in the metal-binding environment (i.e., incorrect side chain rotamers), but they may also indicate incorrect metal identity.

**Identifying the potential causes of errors** ● **TIMING ~10 min**
**16|** Identify potential errors due to the metal-binding environment. The *vacancy*, *geometry* and *nVECSUM* parameters are more influenced by the completeness of the metal-binding environment, as described in Steps 11 and 13, whereas the *gRMSD* and *ligands* parameters are more influenced by problems such as irregularities in the metal-binding environment. Other experimental evidence is usually required to identify the metal identity in such cases, because it is usually impossible to intelligently suggest a more appropriate alternative metal when the environment is poorly defined. However, it may be possible to correct the metal-binding environment first and upload the resulting model to CMM for another analysis of metal identity.
**!** **CAUTION** CMM may suggest alternative metals even if some or all of the five parameters characterizing the binding environment (*vacancy*, *geometry*, *nVECSUM*, *gRMSD* and *ligands*) are problematic; however, in this case, the alternatives it suggests may not be reliable.
**!** **CAUTION** Metal-binding sites in NMR structures usually cannot be determined experimentally by NMR data alone, and assigning a metal ion requires additional information about the identity and locations of metals, either by modeling or by other experiments. Accordingly, alternative metal suggestions in CMM are explicitly disabled for NMR structures. In general, however, CMM will use the same analysis to verify metal sites, regardless of whether the structure was derived from X-ray diffraction or NMR data.

**17|** *Identify potential errors due to misidentification of metals*. Problems of metal identity are less common than problems in the metal-binding environment. The *valence*, *B-factor* and *occupancy* parameters are strong indicators of metal identity. When there are serious problems with one of these three parameters, the metal identity should be carefully reviewed. In certain cases, the *ligands* parameter may also reveal unfavorable metal-ligand interactions. The list of suggested alternative metals should be considered as described below in Step 18B. However, this method has limitations, which are discussed in the 'Experimental design' section.
**?** **TROUBLESHOOTING**

**Correcting problematic metal-binding sites (optional)** ● **TIMING ~up to a few hours**
**18|** Modify the coordinates of the metal-binding site. The chain identifier and residue number of the problematic metal is found in the first column of the CMM results table (**Fig. 3a**). Step 18A below should be followed if the metal-binding

environment is suspicious, and Step 18B should be followed if the identity of the metal may be wrong. Step 18A and 18B are not mutually exclusive, as there may be errors in both environment and metal identity. Start with Step 18A if the cause of the problem is not readily apparent.

(**A**) **Modify the metal-binding environment**

▲ **CRITICAL** Prior experience in molecule modeling and/or refinement is expected to perform the process described in Step 18A. Furthermore, as the details of this procedure depend on the choice of molecular modeling software, the following steps are presented only as a general strategy.

(i) Open the coordinates in an interactive modeling program (e.g., COOT[56]) and center on the metal of interest.

(ii) (X-ray structures only) Open the electron density map. There should be good agreement of the model and electron density.

(iii) Consider flipping the side chains of His, Asn and Gln residues, and check the chirality of small-molecule ligands in the neighboring environment of the metal ion.

(iv) Check for potential atomic clashes in the metal-binding environment by using the MolProbity[57] service (http://molprobity.biochem.duke.edu/).

(v) Refine the new model with a crystallographic refinement program (e.g., REFMAC[18], PHENIX[19] or SHELXL[20]).

(**B**) **Redefine the metal identity**

(i) Modify the atom name, residue name and element name in a text editor. This normally involves substitution of one elemental symbol for another, although the residue name may also depend on the oxidation state. Alternatively, use molecular modeling software to delete an existing metal and place a new one in the same position. If CMM provides a list of suggested alternative metals, CMM can generate PDB files with these metals in place of the current metal. Select the desired metal(s) and click the 'Save' button in the 'Generate a model with alt. metal' box (**Fig. 3d**). Refinement of the new model may not be mandatory, but it is encouraged if resources permit.

▲ **CRITICAL STEP** Consult http://www.wwpdb.org/docs.html for details of the PDB format. Ligand-Expo (http://ligand-expo.rcsb.org/ld-search.html) lists the proper atom names and residue designations for metal ions.

**? TROUBLESHOOTING**

**19|** Return to Step 2 and upload the modified coordinate file. Rebuilding of the environment and/or replacement of the metal should result in improvement of the CMM validation parameters. This process may be iterated multiple times until all CMM parameters fall within a satisfactory range.

## ? TROUBLESHOOTING

Troubleshooting advice can be found in **Table 1**.

**TABLE 1** | Troubleshooting table.

| Step | Problem | Possible reason | Solution |
|---|---|---|---|
| 2A(iii) | CMM server reports the message: 'Invalid PDB file ~ [No such PDBID #### in PDB!]' | The PDB code does not exist in the PDB | Verify that the PDB code exists in the PDB via the PDB website at http://www.rcsb.org |
| 2B(iii) | CMM server reports the message: 'Invalid PDB file ~[###.###: Unspecified error encountered]' | File is not in PDB format, or is not fully compliant with the PDB format | Make sure the uploaded PDB file is fully compliant with the PDB format as specified by http://www.wwpdb.org/docs.html |
| 3 | CMM server reports the message: 'No metal present in the model requested, or metal is far away from the modeled macromolecule chain' | If there is a metal ion in the structure, it is not directly coordinated by any ligand | The server requires at least one metal-ligand interaction for validation. Open the coordinate file in a molecule visualization program and make sure that there is at least one metal-ligand interaction |
| 3 | CMM server does not respond within 1 min | The structure being processed is very large | Some structures may take a longer time to process. For example, a ribosome structure may take up to 15 min to process |
| 5, 7 | The ID button is not shown in the first column of the top result summary table | Java plug-in is not installed, or JavaScript and/or the Java plug-in are disabled | Download and install the Java browser plug-in from http://java.com. Make sure that both JavaScript and Java plug-in are enabled in the web browser |

(continued)

**TABLE 1** | Troubleshooting table. (continued)

| Step | Problem | Possible reason | Solution |
|---|---|---|---|
| 6 | Unable to switch to different metal distribution profiles | JavaScript is disabled | Enable JavaScript in the web browser |
| 7, 8 | Jmol window is not active | The Java plug-in is not installed | Download and install the Java plug-in from http://java.com |
| | Java was blocked because it is out of date | The latest version of Java plug-in is not installed | Download and install the latest version of Java from http://java.com |
| 8, 9 | Rotation of molecule in the Jmol window is not smooth | Poor computer or graphic card performance, and/or insufficient computer memory | Turning off antialiasing can partially alleviate the problem |
| 17 | The list of alternative metals is empty | Alternative metal suggestions are not always possible. Refer to limitations in Experimental design | Identify metal using evidence from biochemical or other experiments<br>Identify metal by manual inspection of the metal-binding environment and geometry<br>Refine metal-binding environment and upload to CMM for another analysis |
| 18B(i) | Cannot locate the three-letter residue code of a metal residue | The metal residue usually has one or two alphabetic characters preceded by two or one empty spaces to form the expected three-letter residue code in the PDB format | The metal residue with one or two alphabetic characters should be right-justified to fit in the three-letter code column in the atom record of the metal atom |



● **TIMING**

Steps 1 and 2, submitting a job: ~10 min
Steps 3 and 4, monitoring the job: ~1 min
Steps 5–9, viewing validation results: ~5 min
Steps 10–15, analyzing problematic sites: ~10 min
Steps 16 and 17, identifying the potential causes of errors: ~10 min
Steps 18 and 19, (optional) correcting problematic metal-binding sites: ~up to a few hours

## ANTICIPATED RESULTS

### All metal-binding sites within the structure are determined to be largely correct (Steps 5–9)

All validation parameters will fall within the acceptable range and be shown in green (or gray when not applicable). In some cases, one or two parameters labeled as borderline (yellow) may be insufficient evidence to characterize a site as problematic. The site may not require further verification or correction, provided that thoughtful consideration is given to those parameters.

### One or more metal-binding site(s) in the structure are problematic (Steps 10–17)

For each problematic metal-binding site, one or more parameter(s) will be displayed in red or yellow, indicating a deviation from regularity. If the optional step of correcting problematic metal-binding sites is not performed or is infeasible, the protocol establishes that the problematic metal-binding site(s) in the structure of interest are most likely unreliable.

### After optional correction (Steps 18 and 19)

All metal-binding sites for the given structure are re-evaluated, and they should show improvement over the initially submitted version.

### Detection and correction of problematic metal-binding sites in X-ray crystallography structures

Four PDB structures determined at good resolution (2 Å or better) with well-characterized function were chosen for the demonstration of specific errors. Two of these examples (PDB codes 3lkm and 2z08) demonstrate the ability of CMM to detect errors in metal assignment when the binding environment is well defined. Two other examples (PDB codes 1qy6 and 3lrk) illustrate that misinterpretation of residual electron density near a metal can lead to an incorrect coordination sphere that

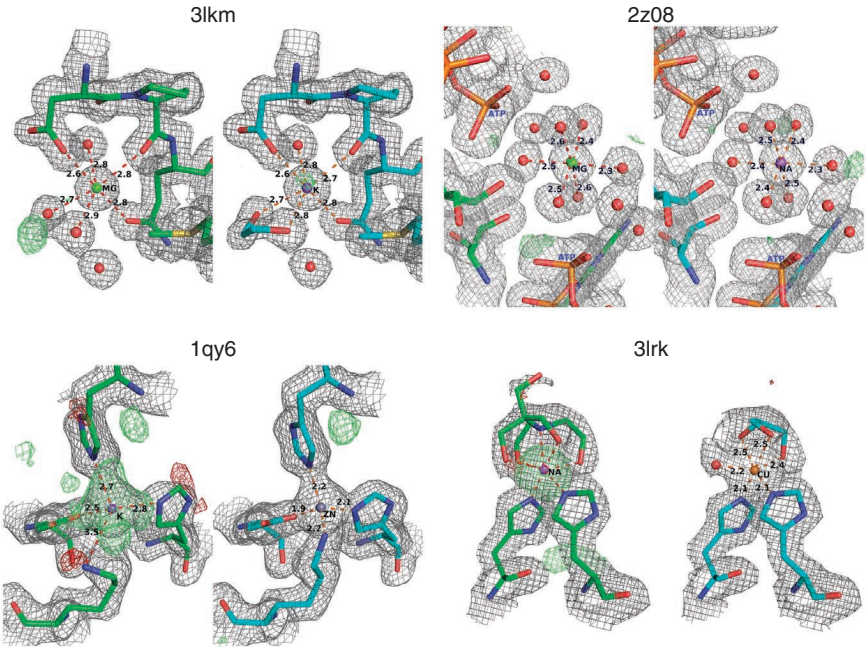



**Figure 4** | Examples of problematic metal-binding sites in crystal structures, as detected by CMM. Metal-binding sites are shown before (protein backbone in green) and after re-refinement (backbone in cyan). Electron density maps ($2F_o$-$F_c$) are shown in gray mesh with a 1.0-σ cutoff, with difference maps ($F_o$-$F_c$) shown in green and red mesh with a 4.0-σ cutoff. Only one of the three modeled magnesium ions (A902) is shown for 3lkm, together with the re-refined site with a potassium ion. Another one of the three modeled magnesium ions in 3lkm, as replaced with water (A901), is shown in **Supplementary Figure 3**. In the 3lrk re-refined structure, only one of the two alternative glycerol molecules is shown for clarity.

should be properly modeled before checking metal identification (**Fig. 4**). By following the re-refinement protocol specified in Step 18A, refinement statistics, model quality (as measured by MolProbity[57]) and the CMM parameters all improved. Refinement statistics for the original PDB file, the optimized model from PDB_REDO[58] and from our re-refinement are shown in **Supplementary Table 3**. The re-refined coordinates (in PDB format) together with structural factor files (in MTZ format) are made available to download via the CMM job submission interface (**Fig. 2**; bottom right, sample coordinates section).

The magnesium-binding sites in the kinase domain of myosin heavy chain kinase A (PDB code 3lkm) have been reported to have an important role for switching the conformation of the active site[59] (**Table 2**). However, CMM indicated that the magnesium ions in this structure are modeled incorrectly. Re-refinement of the structure with the first magnesium ion replaced with water (**Supplementary Fig. 3**) and the other two replaced with potassium ions improves the overall structure quality, including all CMM parameters. Both potassium ions are also coordinated by ethylene glycol, which was used for cryoprotection during sample preparation (**Fig. 4**). The presence of potassium phosphate (0.2 M) in the crystallization buffer (as reported in the PDB record) further justifies this modeling.

CMM indicated an error (a value of 0.8 instead of the expected 2.0 for the *valence* parameter) for one of the two magnesium-binding sites in the structure of an ATP-dependent kinase (PDB code 2z08) with a well-defined binding environment. CMM suggested replacing this magnesium with sodium. Substituting sodium and re-refining with sodium improved the model, resulting in a *valence* value of 1.1 (near the expected value of 1.0). The presence of sodium can be readily explained by the inclusion of 2 M sodium chloride in the published crystallization conditions.

A metal-dependent serine protease (PDB code 1qy6) was modeled with a potassium ion in its active site, coordinated by Asp7, His9 and Lys147 (ref. 60). CMM indicated an error in metal identity. Moreover, examination of the difference map revealed the presence of unexplained electron density. After manually refining the amino acid side chains in the metal-binding environment, CMM suggested replacing the potassium with a transition metal (Zn or Co). Each alternative metal was substituted in turn, and the resulting models were refined individually, and modeling the metal ion as zinc resulted in a better fit according to CMM. Moreover, a homologous protein (PDB code 1wcz) has zinc modeled in the equivalent site.

A model of α-galactosidase (PDB code 3lrk)[61] contains a sodium ion, which CMM reported as improperly modeled with an exceptionally high valence and unusual ligands in its coordination sphere, including the common buffer bis-tris (2-[bis(2-hydroxyethyl)amino]-2-(hydroxymethyl)-1,3-propanediol). The bis-tris molecule and the coordinating sodium were not in agreement with the observed electron density. We also noted an electron density peak of 17σ in the $F_o$-$F_c$ difference map calculated using the original model containing sodium. After replacing bis-tris with alternative conformations of glycerol used for cryoprotection during sample preparation, CMM suggested replacing the sodium with a transition metal (Co or Cu). Alternative models were refined individually, and the model with copper resulted in a slightly better fit. Even though we interpret the replaced metal and its binding environment as more consistent with the electron density data, the proposed glycerol and copper model is not a conclusive solution and needs further experimental verification, such as characteristic copper absorption peaks in X-ray fluorescence spectroscopy[24,25] or copper dependency in an α-galactosidase activity assay. After re-refinement of the model with copper and glycerol, all CMM parameters fell within the acceptable ranges. Knowledge-based function prediction using the extended similarity group (ESG) server[62] calculated a 63% probability for this protein to bind metal cations, although no other published research has indicated that this particular α-galactosidase is metal dependent.

**TABLE 2** | Examples of problematic metal-binding sites in crystal structures as detected by CMM.

| PDB code: residue | Metal | *Occupancy* | B-factor (env.) | *Ligands* | *Valence* (expected) | *nVECSUM* | *Geometry* | *gRMSD* (°) | *Vacancy* |
|---|---|---|---|---|---|---|---|---|---|
| 3lkm: A902 | | | | | | | | | |
| Original | Mg | 1 | *8.0 (13.9)* | **$O_1$**[a] | **0.08 (2.0)** | **1** | **Poorly coordinated** | N/A | N/A |
| Re-refined | K | 1 | 16.7 (16.0) | $O_6$[a] | 1.1 (1.0) | *0.19* | Octahedral | *14.1°* | 0 |
| 2z08: A2002 | | | | | | | | | |
| Original | Mg | **0.5**[b] | *19.1 (24.2)* | $O_6$ | **0.8 (2.0)** | 0.044 | Octahedral | 4.7° | 0 |
| Re-refined | Na | **0.5**[b] | *19.5 (24.2)* | $O_6$ | 1.1 (1.0) | 0.075 | Octahedral | 5.1° | 0 |
| 1qy6: A300 | | | | | | | | | |
| Original | K | 1 | **5.3 (16.4)** | $O_1N_2$ | 0.9 (1.0) | **0.52** | Tetrahedral | *15.4°* | 25% |
| Re-refined | Zn | 1 | 21.8 (21.3) | $O_1N_3$ | *1.6 (2.0)* | *0.19* | Tetrahedral | 8.8° | 0 |
| 3lrk: A8001 | | | | | | | | | |
| Original | Na | 1 | *18.1 (28.4)* | **$O_4N_3$** | **2.3 (1.0)** | *0.12* | **Pentagonal bipyramid** | 11.6° | 0 |
| Re-refined | Cu | 1 | 37.3 (37.9) | $O_4N_2$ | 1.3 (1.0) | *0.19* | *Octahedral* | *13.6°* | 0 |

[a]Only ligands that yield a bond valence value >0.08 are considered as coordinating ligands for purposes of deriving the composition of the *ligands* parameter. This results in a distance cutoff of ~2.65 Å for any oxygen to be considered as a first coordination sphere ligand for magnesium. The distances to the other five oxygens vary between 2.7 and 2.9 Å. At those distances, the oxygen atoms are only considered as coordinating ligands for potassium, not magnesium. [b]The metal is located on a symmetry axis.

The eight parameters are shown before and after re-refinement with text formatting that corresponds to the RAG coding used by the CMM server (**underlined bold** (red in CMM) for outlier, *underlined italics* (yellow in CMM) for borderline, plain text (green in CMM) for acceptable, and **nonunderlined bold** (gray in CMM) for not applicable (N/A)). For structures that contain multiple metal-binding sites, only one metal-binding site is shown for purposes of demonstration. In the re-refined structure of 3lrk, CMM values are obtained using one of the two alternative glycerol molecules coordinating the metal.



These examples show that the presence of one or more outlier values for CMM parameters indicate potential problems in metal-binding site modeling, and re-refinement guided by CMM's hints can lead to an improved model (**Table 2**). Although the improvement of the coordinates of a few atoms will not usually markedly affect global crystallographic parameters, such as the *R* and $R_{free}$ factors, we have observed substantial improvements of statistical values with all four re-refined structures: the crystallographic *R* factor dropped 3–6%, and the $R_{free}$ factor 4–5% (**Supplementary Table 3**). We presume that this is mostly owing to an overall improved refinement strategy. Indeed, the script-driven re-refinement performed by PDB_REDO[58] also improved the *R* and $R_{free}$ statistics without changing the atomic composition.

**Detection of problematic metal-binding sites in NMR structures**

The same set of parameters developed for the validation of metal-binding sites in structures determined by X-ray crystallography also apply for the validation of NMR structures exemplified by several zinc-containing NMR structures (**Table 3**). Nearly ideal zinc-binding sites are found in many NMR structures, including (for example) those defined in various ATRX-DNMT3-DNMT3L domains of the chromatin-associated protein ATRX (PDB code 2lbm) (Eustermann *et al.*[63]). However, problematic zinc-binding sites are not uncommon, such as those in human cancer and autoimmune disease–related E3 protein CBL-B (PDB code 2ldr), which contains two zinc-binding loops[64]. For this example, CMM revealed large deviations of *nVECSUM* and *gRMSD*, indicating problems with metal-binding geometry.

There are also cases where the *geometry* is good but unexpectedly high *valence* values are exhibited. Pluripotency factor Lin28 (PDB code 2li8) possesses a zinc knuckle domain that interacts with the pre-let-7 terminal loop and further inhibits the biogenesis of let-7 miRNAs[65]. Even though it was reported that additional distance constraints of 2.3 Å between the zinc and sulfur ligands were used, CMM indicated that several metal-ligand distances were significantly shorter (around 2.1 Å) in the refined model, and thus determined an unexpectedly high *valence* value of 3.3 instead of the expected 2.0. Another example of a structure with unexpectedly high *valence* values is the archaeal endonuclease Nob1 (PDB code 2lcq), which has a zinc ribbon domain that is common in many nucleases[66]. An initial zinc-sulfur distance of 2.3 Å was used with distance restraints of 2.1–2.6 Å applied between the zinc atom and the sulfur atoms of the four coordinating cysteine residues.

**TABLE 3** | Examples of plausible and problematic zinc-binding sites in NMR structures as detected by CMM.

| PDB code: residue | Metal | *Ligands* | *Valence* (expected) | *nVECSUM* | *Geometry* | *gRMSD* (°) | *Vacancy* |
|---|---|---|---|---|---|---|---|
| 2lbm: A1 | Zn | $S_4$ | 2.2 (2.0) | 0.03 | Tetrahedral | 2.9° | 0 |
| 2ldr: A1373 | Zn | $S_4$ | 1.7 (2.0) | **0.6** | Tetrahedral | **35.7°** | 0 |
| 2li8: A187 | Zn | $N_1S_3$ | **3.3 (2.0)** | 0.1 | Tetrahedral | 5.6° | 0 |
| 2lcq: A162 | Zn | $S_4$ | **4.6 (2.0)** | 0.033 | Tetrahedral | 6.1° | 0 |
| 2fuu: A201 | Zn | $N_1S_2$ | **14.4 (2.0)** | **0.93** | *Octahedral* | **24.9°** | **50%** |

The six parameters are shown for each metal site with text formatting that corresponds to the RAG coding used on the CMM server (**underlined bold** (red in CMM) for outlier, *underlined italics* (yellow in CMM) for borderline, plain text (green in CMM) for acceptable). Zinc-binding sites across different models of the ensemble are of similar quality and the parameters of only the first model in each ensemble is shown for purposes of demonstration. Similarly, for structures that contain more than one zinc-binding site, they are of similar quality and only one of them is shown.

However, the final models have zinc-binding sites with zinc-sulfur distances consistently below 2.1 Å, resulting in an even higher *valence* value of 4.6 instead of the expected 2.0 (**Table 3**).

In yet another example, CMM indicated that both *valence* and *geometry* deviate markedly from the acceptable range for a bromodomain PHD finger transcription factor (PDB code 2fuu), which contains two compact zinc-finger scaffolds[67]. A crystal structure of the same protein (PDB code 2f6j) at 2.0-Å resolution was also reported in the same study, but with both zinc-binding sites properly defined. These representative examples illustrate a variety of pitfalls that one might encounter during metal-binding site modeling in NMR structures. The same techniques can be easily applied to other metals.

*Note: Any Supplementary Information and Source Data files are available in the online version of the paper.*

**ACKNOWLEDGMENTS** This work was supported by Federal funds from the National Institute of Allergy and Infectious Diseases, US National Institutes of Health, Department of Health and Human Services, under contract nos. HHSN272200700058C and HHSN272201200026C. We thank M. Grabowski, K.M. Langner and M. Domagalski for the CSGID website framework containing the CMM server; J. Hou, I.G. Shabalin, I.A. Shumilin, M. Demas and A.A. Knapik for server testing; W.F. Anderson for valuable discussion; and M.D. Zimmerman and H.C. Chapman for critically reading the manuscript.

**AUTHOR CONTRIBUTIONS** H.Z. designed, implemented, tested and maintained the CMM server; H.Z. developed, implemented and optimized the NEIGHBORHOOD database; M.D.C. and D.R.C. helped design the geometry assignment algorithm and server interface; D.R.C. implemented the first version of the Jmol applet; P.M. and G.M.S. introduced the CBVS and VECSUM methods, which were slightly modified for this study; H.Z., M.D.C., D.R.C., M.C., P.M., G.M.S. and W.M. wrote and approved the manuscript; and W.M. supervised the project.

**COMPETING FINANCIAL INTERESTS** The authors declare no competing financial interests.



1. Harding, M.M., Nowicki, M.W. & Walkinshaw, M.D. Metals in protein structures: a review of their principal features. *Crystallogr. Rev.* **16**, 247–302 (2010).
2. Berman, H.M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
3. Pozharski, E., Weichenberger, C.X. & Rupp, B. Techniques, tools and best practices for ligand electron-density analysis and results from their application to deposited crystal structures. *Acta Crystallogr. D* **69**, 150–167 (2013).
4. Chruszcz, M., Domagalski, M., Osinski, T., Wlodawer, A. & Minor, W. Unmet challenges of structural genomics. *Curr. Opin. Struct. Biol.* **20**, 587–597 (2010).
5. Zheng, H., Chruszcz, M., Lasota, P., Lebioda, L. & Minor, W. Data mining of metal ion environments present in protein structures. *J. Inorg. Biochem.* **102**, 1765–1776 (2008).
6. Branden, C. & Jones, T. Between objectivity and subjectivity. *Nature* **343**, 687–689 (1990).
7. Adams, P.D. *et al.* Advances, interactions, and future developments in the CNS, Phenix, and Rosetta structural biology software systems. *Annu. Rev. Biophys.* **42**, 265–287 (2013).
8. Minor, W., Cymborowski, M., Otwinowski, Z. & Chruszcz, M. HKL-3000: the integration of data reduction and structure solution—from diffraction images to an initial model in minutes. *Acta Crystallogr. D* **62**, 859–866 (2006).
9. Chen, V.B. *et al.* MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D* **66**, 12–21 (2010).
10. Abriata, L.A. Investigation of non-corrin cobalt(II)-containing sites in protein structures of the Protein Data Bank. *Acta Crystallogr. B* **69**, 176–183 (2013).
11. Dauter, Z., Weiss, M.S., Einspahr, H. & Baker, E.N. Expectation bias and information content. *Acta Crystallogr. F* **69**, 83 (2013).
12. Weichenberger, C.X., Pozharski, E. & Rupp, B. Visualizing ligand molecules in Twilight electron density. *Acta Crystallogr. F* **69**, 195–200 (2013).
13. Wlodawer, A., Minor, W., Dauter, Z. & Jaskolski, M. Protein crystallography for non-crystallographers, or how to get the best (but not more) from published macromolecular structures. *FEBS J.* **275**, 1–21 (2008).
14. Nayal, M. & Di Cera, E. Valence screening of water in protein crystals reveals potential $Na^+$ binding sites. *J. Mol. Biol.* **256**, 228–234 (1996).
15. Nabuurs, S.B., Spronk, C.A., Vuister, G.W. & Vriend, G. Traditional biomolecular structure determination by NMR spectroscopy allows for major errors. *PLoS Comput. Biol.* **2**, e9 (2006).
16. Hsin, K., Sheng, Y., Harding, M.M., Taylor, P. & Walkinshaw, M.D. MESPEUS: a database of the geometry of metal sites in proteins. *J. Appl. Crystallogr.* **41**, 963–968 (2008).
17. Abriata, L.A. Analysis of copper-ligand bond lengths in X-ray structures of different types of copper sites in proteins. *Acta Crystallogr. D* **68**, 1223–1231 (2012).
18. Murshudov, G.N. *et al.* REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallogr. D* **67**, 355–367 (2011).
19. Adams, P.D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66**, 213–221 (2010).
20. Sheldrick, G.M. A short history of SHELX. *Acta Crystallogr. A* **64**, 112–122 (2008).
21. Bergerhoff, G. & Brandenburg, K. in *International Tables for Crystallography* (eds. Wilson, J.C. & Prince, E.) 778–789 (John Wiley & Sons, 2006).
22. Laskowski, R.A., MacArthur, M.W., Moss, D.S. & Thornton, J.M. PROCHECK—a program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.* **26**, 283–291 (1993).
23. Vaguine, A.A., Richelle, J. & Wodak, S.J. SFCHECK: a unified set of procedures for evaluating the quality of macromolecular structure-factor data and their agreement with the atomic model. *Acta Crystallogr. D* **55**, 191–205 (1999).
24. Ascone, I. & Strange, R. Biological X-ray absorption spectroscopy and metalloproteomics. *J. Synchrotron Radiat.* **16**, 413–421 (2009).

25. Garcia, J.S., Magalhaes, C.S. & Arruda, M.A. Trends in metal-binding and metalloprotein analysis. *Talanta* **69**, 1–15 (2006).
26. Müller, P., Köpke, S. & Sheldrick, G.M. Is the bond-valence method able to identify metal atoms in protein structures? *Acta Crystallogr. D* **59**, 32–37 (2003).
27. Tylichova, M. *et al.* Structural and functional characterization of plant aminoaldehyde dehydrogenase from *Pisum sativum* with a broad specificity for natural and synthetic aminoaldehydes. *J. Mol. Biol.* **396**, 870–882 (2010).
28. Seff, A.L., Pilbak, S., Silaghi-Dumitrescu, I. & Poppe, L. Computational investigation of the histidine ammonia-lyase reaction: a modified loop conformation and the role of the zinc(II) ion. *J. Mol. Model.* **17**, 1551–1563 (2011).
29. Srikanth, R., Mendoza, V.L., Bridgewater, J.D., Zhang, G. & Vachet, R.W. Copper binding to β-2-microglobulin and its pre-amyloid oligomers. *Biochemistry* **48**, 9871–9881 (2009).
30. Cooper, D.R., Porebski, P.J., Chruszcz, M. & Minor, W. X-ray crystallography: assessment and validation of protein-small molecule complexes for drug discovery. *Exp. Opin. Drug Discov.* **6**, 771–782 (2011).
31. Pietrzyk, A.J. *et al.* High-resolution structure of *Bombyx mori* lipoprotein 7: crystallographic determination of the identity of the protein and its potential role in detoxification. *Acta Crystallogr. D* **68**, 1140–1151 (2012).
32. Brown, I.D. Recent developments in the methods and applications of the bond valence model. *Chem. Rev.* **109**, 6858–6919 (2009).
33. Hanson, R.M. Jmol—a paradigm shift in crystallographic visualization. *J. Appl. Crystallogr.* **43**, 1250–1260 (2010).
34. Allen, F.H. The Cambridge Structural Database: a quarter of a million crystal structures and rising. *Acta Crystallogr. B* **58**, 380–388 (2002).
35. Brylinski, M. & Skolnick, J. FINDSITE-metal: integrating evolutionary information and machine learning for structure-based metal-binding site prediction at the proteome level. *Proteins* **79**, 735–751 (2011).
36. Sodhi, J.S. *et al.* Predicting metal-binding site residues in low-resolution structural models. *J. Mol. Biol.* **342**, 307–320 (2004).
37. Cai, C.Z., Han, L.Y., Ji, Z.L., Chen, X. & Chen, Y.Z. SVM-Prot: Web-based support vector machine software for functional classification of a protein from its primary sequence. *Nucleic Acids Res.* **31**, 3692–3697 (2003).
38. Levy, R., Edelman, M. & Sobolev, V. Prediction of 3D metal binding sites from translated gene sequences based on remote-homology templates. *Proteins* **76**, 365–374 (2009).
39. Passerini, A., Lippi, M. & Frasconi, P. MetalDetector v2.0: predicting the geometry of metal binding sites from protein sequence. *Nucleic Acids Res.* **39**, W288–W292 (2011).
40. Hemavathi, K. *et al.* MIPS: metal interactions in protein structures. *J. Appl. Crystallogr.* **43**, 196–199 (2010).
41. Castagnetto, J.M. *et al.* MDB: the Metalloprotein Database and Browser at The Scripps Research Institute. *Nucleic Acids Res.* **30**, 379–382 (2002).
42. Andreini, C., Cavallaro, G., Lorenzini, S. & Rosato, A. MetalPDB: a database of metal sites in biological macromolecular structures. *Nucleic Acids Res.* **41**, D312–D319 (2013).
43. Andreini, C., Bertini, I., Cavallaro, G., Holliday, G.L. & Thornton, J.M. Metal-MACiE: a database of metals involved in biological catalysis. *Bioinformatics* **25**, 2088–2089 (2009).
44. Degtyarenko, K.N., North, A.C. & Findlay, J.B. PROMISE: a database of bioinorganic motifs. *Nucleic Acids Res.* **27**, 233–236 (1999).
45. Laskowski, R.A. PDBsum new things. *Nucleic Acids Res.* **37**, D355–D359 (2009).
46. Golovin, A. & Henrick, K. MSDmotif: exploring protein sites and motifs. *BMC Bioinformatics* **9**, 312 (2008).
47. Pettersen, E.F. *et al.* UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).
48. Brese, N.E. & O'Keeffe, M. Bond-valence parameters for solids. *Acta Crystallogr. B* **47**, 192–197 (1991).
49. Shields, G.P., Raithby, P.R., Allen, F.H. & Motherwell, W.D. The assignment and validation of metal oxidation states in the Cambridge Structural Database. *Acta Crystallogr. B* **56** (Part 3): 455–465 (2000).
50. Carugo, O. & Djinovic Carugo, K. When X-rays modify the protein structure: radiation damage at work. *Trends Biochem. Sci.* **30**, 213–219 (2005).
51. Hersleth, H.P. & Andersson, K.K. How different oxidation states of crystalline myoglobin are influenced by X-rays. *Biochim. Biophys. Acta* **1814**, 785–796 (2011).
52. Katz, A., Glusker, J., Beebe, S. & Bock, C. Calcium ion coordination: A comparison with that of beryllium, magnesium, and zinc. *J. Am. Chem. Soc.* **118**, 5752–5763 (1996).
53. Harding, M.M. The architecture of metal coordination groups in proteins. *Acta Crystallogr. D* **60**, 849–859 (2004).
54. Kuppuraj, G., Dudev, M. & Lim, C. Factors governing metal-ligand distances and coordination geometries of metal complexes. *J. Phys. Chem. B* **113**, 2952–2960 (2009).
55. Bailey, S. The CCP4 suite—programs for protein crystallography. *Acta Crystallogr. D* **50**, 760–763 (1994).
56. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60**, 2126–2132 (2004).
57. Lovell, S.C. *et al.* Structure validation by Cαgeometry: ϕ, ψ and Cβ deviation. *Proteins* **50**, 437–450 (2003).
58. Joosten, R.P., Joosten, K., Cohen, S.X., Vriend, G. & Perrakis, A. Automatic rebuilding and optimization of crystallographic structures in the Protein Data Bank. *Bioinformatics* **27**, 3392–3398 (2011).
59. Ye, Q., Crawley, S.W., Yang, Y., Cote, G.P. & Jia, Z. Crystal structure of the α-kinase domain of *Dictyostelium* myosin heavy chain kinase A. *Sci. Signal.* **3**, ra17 (2010).
60. Prasad, L., Leduc, Y., Hayakawa, K. & Delbaere, L.T. The structure of a universally employed enzyme: V8 protease from *Staphylococcus aureus*. *Acta Crystallogr. D* **60**, 256–259 (2004).
61. Yoshiba, S. *et al.* Structural insights into the *Thermus thermophilus* ADP-ribose pyrophosphatase mechanism via crystal structures with the bound substrate and metal. *J. Biol. Chem.* **279**, 37163–37174 (2004).
62. Chitale, M., Hawkins, T., Park, C. & Kihara, D. ESG: extended similarity group method for automated protein function prediction. *Bioinformatics* **25**, 1739–1745 (2009).
63. Eustermann, S. *et al.* Combinatorial readout of histone H3 modifications specifies localization of ATRX to heterochromatin. *Nat. Struct. Mol. Biol.* **18**, 777–782 (2011).
64. Kobashigawa, Y. *et al.* Autoinhibition and phosphorylation-induced activation mechanisms of human cancer and autoimmune disease-related E3 protein Cbl-b. *Proc. Natl. Acad. Sci. USA* **108**, 20579–20584 (2011).
65. Loughlin, F.E. *et al.* Structural basis of pre-let-7 miRNA recognition by the zinc knuckles of pluripotency factor Lin28. *Nat. Struct. Mol. Biol.* **19**, 84–89 (2011).
66. Veith, T. *et al.* Structural and functional analysis of the archaeal endonuclease Nob1. *Nucleic Acids Res.* **40**, 3259–3274 (2011).
67. Li, H. *et al.* Molecular basis for site-specific read-out of histone H3K4me3 by the BPTF PHD finger of NURF. *Nature* **442**, 91–95 (2006).