Automated oxidation-state assignment for metal sites in coordination complexes in the Cambridge Structural Database

STRUCTURAL SCIENCE CRYSTAL ENGINEERING MATERIALS

ISSN 2052-5206

Automated oxidation-state assignment for metal sites in coordination complexes in the Cambridge Structural Database

Matthew G. Reeves,[a] Peter A. Wood[b]* and Simon Parsons[a]*

[a]Centre for Science at Extreme Conditions and EaStCHEM School of Chemistry, The University of Edinburgh, King's Buildings, West Mains Road, Edinburgh EH9 3FJ, Scotland, and [b]Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, UK. *Correspondence e-mail: wood@ccdc.cam.ac.uk, s.parsons@ed.ac.uk

Received 9 August 2019
Accepted 21 September 2019

Edited by M. Du, Tianjin Normal University, People's Republic of China

Keywords: Cambridge Structural Database; coordination chemistry; oxidation states; transition metals; bond-valence sum method.

Supporting information: this article has supporting information at journals.iucr.org/b

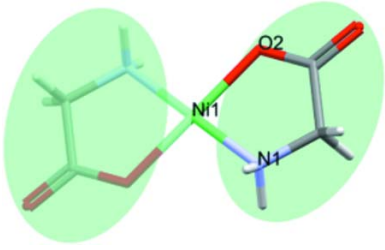
The Cambridge Structural Database (CSD) currently contains over 400 000 transition-metal-containing entries, however many entries still lack curated oxidation-state assignments. Surveying and editing the remaining entries would be far too resource- and time-intensive to be carried out manually. Here, a highly reliable automated workflow for oxidation-state assignment in transition-metal coordination complexes via CSD Python API (application programming interface) scripts is presented. The strengths and limitations of the bond-valence sum (BVS) method are discussed and the use of complementary methods for improved assignment confidence is explored. In total, four complementary techniques have been implemented in this study. The resulting workflow overcomes the limitations of the BVS approach, widening the applicability of an automated procedure to more CSD entries. Assignments are successful for 99% of the cases where a high consensus between different methodologies is observed. Out of a total number of 54 999 unique metal atoms in a test dataset, the procedure yielded the correct oxidation state in 47 072 (86%) of cases.

1. Introduction

The Cambridge Structural Database (CSD; Groom *et al.*, 2016) currently contains over 400 000 structures of coordination complexes but only about half of these entries specify the metal oxidation state and this is reported in the compound name field. Although the current system of incorporating oxidation states in the entry compound name provides some scope for filtering entries, it would be much more advantageous to associate specific oxidation states with individual transition-metal sites. In this way it would be possible to distinguish specific oxidation states in polynuclear complexes as well as quickly filter entries by both metal and associated valence.

As the CSD moves to a new data structure which includes oxidation state as an atomic property, new processes are needed to generate and assign individual valences. Given the number of transition-metal-containing entries, it would be impractical to attempt manual identification and curation of current, as well as future, entries. Automated processes that can distinguish individual atomic valences are therefore highly desirable.

Past studies have identified and validated transition-metal oxidation states using a combination of bond-valence sum (BVS) and ligand-templating processes. Shields *et al.* (2000) implemented a two-step process in oxidation-state assignment



and validation. An initial oxidation state was estimated using a ligand-templating method whereby an algorithm was trained to recognize common ligand templates surrounding a metal centre and then apply the charge associated with each ligand to determine the corresponding charge of the metal. Having achieved this, BVS was applied to the structure using parameters associated with the oxidation state interpreted from the results of the ligand-template method.

This method was applied to a subset of 743 manually verified copper +1 and +2 structures, with 98% successful assignment. While these results are extremely promising, the procedure relied on appropriate coverage of ligand templates and BVS parameters to produce a confident assignment. Where either the template or BVS method deviated from the expected value, manual inspection was required to check the assignments made.

The BVS method has also been applied to inorganic compounds in the Inorganic Crystal Structure Database with the aim of predicting the formation of likely oxidation states in the presence of specific anions (Davies *et al.*, 2018).

Here we present new methods that can be broadly applied to molecular coordination complexes, in most cases without the need for manual intervention. As in previous work, oxidation states are assigned using the BVS method but without the need to assume or derive an initial oxidation-state estimate. The BVS calculations are supported by the assignment of ligand charges but avoid the definition of templates (in most cases, open-shell ligands are an exception). The combination of these methods provides a confidence-scored oxidation state for each metal atom present in a complex. All calculations make use of the CSD Python API (application programming interface), which has been distributed alongside the CSD since 2015. A stand-alone script intended for use with individual cifs is also available.

2. Methodology

2.1. Assignment of oxidation states using the bond-valence method

In the bond-valence method each metal(i)–ligand(j) bond is assigned a valence, S_{ij}, based on its length and two empirical parameters, R_0 and B. The sum of the valences of the bonds formed by the metal is its oxidation state (Brown, 2016*a*). Bond-valence parameters depend on the metal, its oxidation state and the identity of the bonded ligand atom.

The parameters R_0 and B are taken from the database compiled by Brown (2016*b*) and bonds are defined using the default CSD chemical connectivity cut-offs. The calculation is carried out for all available bond-valence parameters provided that, for each metal–ligand bond in the molecule, parameters exist for all common oxidation states. An oxidation state was considered common if it applied to more than 15% of a metal's assigned entries in the CSD [a list is given in the supporting information; this choice of cut-off gives a listing broadly similar to that in Housecroft and Sharpe's popular textbook (Housecroft & Sharpe, 2008)]. The value of S_{ij} is calculated for each of the oxidation states for which values of R_0 and B are available in Brown's database [equation (1), where R_{ij} is the metal–ligand bond distance].

$$S_{ij} = \exp\left(\frac{R_0 - R_{ij}}{B}\right), \tag{1}$$

For example, the chromium compound HIQYAJ (Chérif *et al.*, 2013) contains the $[Cr(oxalate)_2(H_2O)_2]^-$ anion. The common oxidation states for Cr are +2 and +3, so unless parameters for Cr—O bonds for both are available no attempt is made to assign the oxidation state at all. In fact parameters are available for Cr—O bonds in all oxidation states from Cr(+2) to Cr(+6) and all of these are considered in the oxidation-state assignment procedure.

For each oxidation state, the values of S_{ij} are summed to give a total trial oxidation state, V_t. The value of BVS is compared with the oxidation state (V_p) corresponding to the bond-valence parameters used to calculate it.

$$\Delta = |V_t - V_p|. \tag{2}$$

The oxidation state of the metal is taken as the one which yields the smallest value of Δ, that is, the oxidation state which is most consistent with the parameters used to calculate it.

For example, in the four coordinate cobalt complex KUYHES (Akbarzadeh Torbati *et al.*, 2010) there are two Co–N bonds with distances 2.042 and 2.053 Å and two Co–Cl bonds measuring 2.219 and 2.217 Å. Cobalt has two common oxidation states, +2 and +3, and Co–N and Co–Cl bond-valence parameters are available for both. Using the Co(+2)–N and Co(+2)–Cl bond-valence parameters to calculate the bond valences of the Co–N and Co–Cl bonds yields a total trial valence (V_t) of 1.987. The difference, Δ, between this and the reference oxidation state used to select the bond-valence parameters (V_p) is $|1.987 - 2| = 0.013$. Using the bond-valence parameters for Co(+3)–N and Co(+3)–Cl bonds yields V_t = 2.014 and $\Delta = |2.014 - 3| = 0.986$. Since $0.013 < 0.986$, the oxidation state of the cobalt is taken as +2.

If the minimum value of Δ is greater than 0.5 a warning is added to the assignment. Warnings are used in confidence scoring (see Section 2.3).

For many bond types, R_0 and B have been determined multiple times. Different parameters may apply to different spin states, *e.g.* high and low spin Fe(+2)–N, or be derived from different classes of compound or datasets of different sizes. Each available set of parameters was used to calculate a value of Δ, with the smallest value being used to assign the oxidation state of the metal. The procedure, which was found to reproduce known oxidation states more reliably than using the parameters designated 'most reliable' in Brown's database, is explained in detail in the supporting information for the KUYHES example of the previous paragraph.

2.2. Assignment of oxidation states using ligand charges

2.2.1. Ligand-charge assignment procedure. As an alternative to the bond-valence method, likely ligand charges were also determined using the very fast semi-empirical electronic

structure package *MOPAC* (Stewart, 2016). The overall charge on a complex can be derived from the sum of the formal atomic charges that are stored in the CSD as atomic charge properties. Therefore, the metal oxidation state can be assigned as the sum of the stored formal atomic charges minus the sum of the ligand charges (see §§ 2.2.2, 2.2.3 and 2.2.4).

In the first stage of the procedure the metal centre is removed, leaving the ligand fragments for charge assignment. This process only considers ligands directly connected to the metal centre of interest. Where a salt occurs in the database, formal atomic charges are added to the metal centre by the scientific editors at the Cambridge Crystallographic Data Centre (CCDC) to achieve a charge-neutral structure.

For each ligand fragment the total number of electrons is determined assuming charge neutrality. For ligands with an even number of electrons, possible charges were taken to be +4 to −8 in steps of 2e; for those with an odd number of electrons, possible charges were +3 to −9, also in steps of 2e. This procedure does not consider the possibility that a ligand has an open-shell (*i.e.* a radical) electron configuration. Radicals are discussed below along with further comments on cationic ligands. For each charge, a single-point electronic structure calculation (*MOPAC*) is carried out using the crystal structure geometry and the PM7 method (Stewart, 2013). Each calculation yields a heat of formation and a Parr and Pople hardness parameter. The charge is assigned on the basis of the formation energy and the hardness parameter.

2.2.2. Charge assignment using the Parr & Pople hardness parameter. The hardness parameter quantifies the resistance to changes in the electron configuration (Pearson, 1993) and the charge on the ligand was taken as the one yielding the largest hardness parameter. Hardness values typically fall into the range 0–10 eV. Any charges yielding a hardness outside this range are disregarded. A warning is issued if the difference in hardness is <1 eV.

2.2.3. Charge assignment using formation energy. The charge on the ligand was taken as the one yielding the most negative formation energy. As an example, the energy versus charge plot for a ligand with formula NO_3 is shown in Fig. 1. There is a clear minimum for a charge of −1, indicating that the ligand is NO_3^-.

For some structures, a ligand fragment may produce a set of formation energies with a shallow minimum, making charge assignment ambiguous. Experience showed that ambiguities arose when the energy difference between charges was lower than 150 kJ mol^{-1}. Values lower than this could, for example, lead to assignment of different charges for identical ligands in different crystal structures. Where energy differences do suggest a shallow minimum, a warning is added to the fragment assignment and this is carried forward when considering overall assignment confidence.

2.2.4. Assignment of oxidation states using hydrogen-placement algorithms. The CSD Python API has a built-in molecular editing tool for automatic hydrogen placement which can be used to determine the charge of the ligand following removal of the metal atoms (as in Section 2.2.1). The number of H atoms in a ligand is first recorded. The H atoms

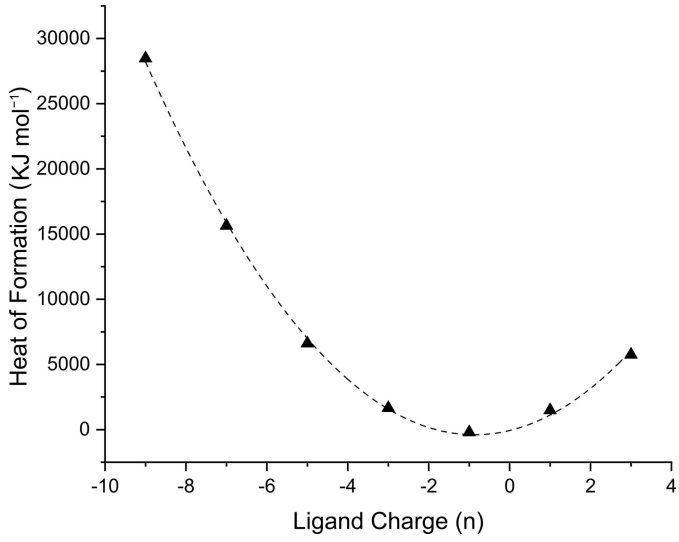


Figure 1
Heat of formation (kJ mol^{-1}) versus ligand charge (n) for NO_3^n. Energies calculated in single-point energy calculations in the crystal structure geometry using *MOPAC*.

are all removed and then replaced using the H-atom generation routine assuming charge neutrality. The difference between the number of H atoms before and after this procedure is the charge. For example, the methoxide ligand CH_3O^- contains three H atoms. Removal of these followed by automatic H-atom placement generates methanol, CH_3OH, containing four H atoms. The charge on the original CH_3O fragment is therefore −1 since the addition of one proton is required to generate a neutral molecule. Having determined the ligand charges in this way the oxidation state of the metal is assigned following the procedure of Section 2.2.1.

2.2.5. Radicals. The ligand-charge calculation is carried out in steps of 2e because the problems associated with shallow minima become much more common if charges are sampled in steps of 1e. The number of structures containing radical ligands is quite small, <2% of structures in the test set used work for method validation (Section 3.1). For common radical species, these ligands can be identified beforehand from their SMILES formulae and are added manually to an SQLite database (Hipp, 2019) in the form of an exceptions list look-up table, which pre-assigns a ligand fragment charge before any determination processes are carried out (see Section 3.2.5). This procedure is similar to the templating method used by Shields *et al.* in their work.

2.2.6. Cationic ligands. Cations can be readily identified in entries from the CSD by the systematic presence of positive symbols in the SMILES formulae generated by the CSD Python API, so that the charge of the fragment is determined by simply summing the number of positive symbols and subtracting the number of negative symbols within the SMILES formula. As SMILES-based charge labelling requires specifically ionic atomic sites, this process cannot be used to distinguish between neutral and anionic ligands, where the metal–ligand bond is typically considered as neutral.

Where a ligand is zwitterionic, there is an ambiguity as to how ligands have been labelled. In the Zn^{2+} complex CSD refcode EGAPOR (Torzilli *et al.*, 2002), zwitterionic *N*-*n*-propylsalicylaldimine-*O* ligands are identified in the CSD entry, with ligating atoms denoted with a negative charge and the protonated imine nitrogen atom with a positive charge. By contrast, in the Cu(+2)-containing refcode CICWIU (Rotondo *et al.*, 1984) only the positive charge on a terminal ammonium moiety is identified in the SMILES formula N(=C\c1ccccc1O)/CCNCCNCC[NH3+], which suggests a +1 cation rather than the true neutral overall ligand. In order to address these issues, assignments made using this method are only accepted providing the closed-shell requirement described earlier is still obeyed. Where this is not the case, a warning is displayed and the corresponding assignment of metal oxidation state is aborted.

2.2.7. Oxidation-state assignment based on ligand charges. The metal oxidation state is determined for mononuclear complexes from the total charge of the ligands and the overall charge on the complex to achieve a net-neutral crystal structure.

The same approach can be applied to polynuclear complexes where there is a single metal atom in the asymmetric unit and assuming charge order so that the overall charge is split evenly between each metal atom present in the overall structure. As an example, the dimeric Cu complex SAVRIQ01 (Mezei & Raptis, 2004) is located on an inversion centre so that the asymmetric unit contains a single copper atom. Assigning all ligand charges in the complex gives an overall charge of +4. Using the assumption that each asymmetric unit has the same valence, the valence of each copper atom is equal to $\frac{1}{2}$ (*i.e.* $1/n$ asymmetric units that make up a complete molecule) times the overall charge = +2.

In other polynuclear complexes the total ligand charge can only be used to obtain the sum of the metal oxidation states, and BVS is the only method capable of assigning the oxidation states of individual metal atoms. The total ligand charge is used instead for validating the BVS assignments.

2.3. Oxidation-state assignment and confidence scoring

The preceding sections have described four methods for oxidation-state assignment: a BVS approach and three ligand charge-based methods using minimum energy, maximum hardness and the number of H atoms. It is only strictly necessary to apply these methods to a new CSD entry in cases where an author-supplied oxidation state is not available, though we recommend that they could also be used to validate author assignments.

Where named valences are not available, assignments are made using a combination of all the methods described. In ideal cases, all four methods should agree on the assigned oxidation state. In cases where the methods disagree, the oxidation-state assignment is attempted using the BVS method as this is the only method that can be applied to both the mononuclear and polynuclear complexes. Where BVS cannot be applied, the assignment is made based on the maximum hardness method for ligand-charge assignment. The reliability of this method is very similar to the energy-assignment method but during testing there were found to be fewer ambiguous cases (see above) than for the energy method.

Table 1
Confidence-scoring values for each assignment method.

Scores are given for agreement with the most reliable method.

Method	Method agrees with assignment (without errors/warnings)
BVS	5 (+1)
MOPAC ligand assignments by hardness	4 (+1)
MOPAC ligand assignments by energy	3 (+1)
Ligand assignments by hydrogen placement	2

Table 2
Confidence-score grade bandings.

The reliability is based on the results of Section 3.1.

Score	Band values	Description	Reliability (%)
0	U	Unassigned	0
0–5	D	Very unreliable	18
6–8	C	Quite unreliable	56
9–12	B	Reliable	98
>12	A	Very reliable	>99

Following oxidation-state assignment, a confidence score is determined based on the success rate of each method, the agreement between different methods and the occurrence of any warnings. A numeric score is determined using a summation of values from Table 1. Each assignment may have an overall score between 0 (no assignment) and 17 (all assignments agree without error). For simplicity, these are banded into letter grades (A–D) as in Table 2, with A indicating the highest level of confidence and D indicating the lowest level of confidence.

The confidence bands have been defined on the basis of experience, based on which methods were most effective at predicting the correct oxidation state. Examples are given below.

2.4. Ligand database

The three charge-assignment methods described above have been used to generate an SQLite database of ligands, their frequency in the CSD and the assigned charge. The database can be accessed, updated and added to through the SQLite3 Python module (Hipp, 2019).

The database contains the SMILES formula for each ligand in the CSD and the number of times it has been encountered. The number of entries is currently 12 939. Most ligands appear in multiple CSD entries, yielding a distribution of charges for each of the three methods described in §§2.2.2, 2.2.3 and 2.2.4. For each ligand the modes (*i.e.* the most common values) and standard deviations of the each of the three distributions are stored in the database. These data enable a proposed charge assignment to be checked against previous assignments for the

same ligand, while also providing a measure of confidence in the comparison.

The database facilitates the ability to override potentially incorrect charge assignments where the value disagrees with previous values. In order to achieve this, for each ligand fragment encountered, the mode of previous assignments is compared with the currently determined value. If the value does not match the most common charge determined in the database, the database value is used instead, with a warning generated reducing the confidence score by one.

2.5. Confidence-scoring examples

Example 1. In the entry AMIRAR (Holler *et al.*, 2016) where Cu(+1) is coordinated to an acetonitrile and a thiopyridazine scorpionate ligand in which both the N and B atoms are bound to the Cu, the BVS method could not be applied because Cu—B parameters are unavailable. As a result, only the ligand-charge methods can be applied. Application of the minimum-energy method yields an oxidation state of +3, the hardness method a value of +1 and the H-atom placement method a value of +1; no errors were issued in any of these procedures. The value selected is taken from the hardness method, correctly assigning a valence of +1. The confidence score is the sum of 0 for the BVS method, 5 for the hardness method (4 + 1 for no errors in assignment), 0 for the energy since this disagrees with the results from the method with highest reliability and 2 for the hydrogen-placement method. For both ligands there are previous assignments available
in the SQLite database and for all methods the mode charge agrees with the current assignment. The total score is 7; this C-grade assignment should be considered quite unreliable.

Example 2. In chlorobis(*N*-phenylbenzohydroxamato)(triphenylphosphine)rhodium(3+) (refcode CAFSEI; Das *et al.*, 2002), BVS assignment is not possible because of a lack of Rh–P/Rh–Cl parameters and assignment must be made using ligand-charge methods only. For the chloride and triphenylphosphine ligands, the hardness method correctly assigns charges of −1 and 0, respectively. However, the *N*-phenylbenzohydroxamato ligands are incorrectly given a charge of +1. This ligand is listed in the ligand database with nine previous assignments, with a (correct) mode of −1. The database value charge (−1) replaces that determined by the hardness method and a warning is associated with the hardness method. The hardness confidence score is therefore 5 − 1 = 4. The energy and hydrogen-placement methods both yield the correct charge of −1, so all three methods produce the correct valence of Rh(+3) for this structure, the final confidence therefore is the sum of the scores for BVS (0), energy (4), hardness (4) and hydrogen placement (2) = 10, lying in the B confidence band. This B-grade assignment should be considered reliable.

3. Discussion

3.1. Success rate of oxidation-state assignment

The aim of the present study was to determine the oxidation states of transition metals in coordination complexes using an automated procedure and to devise a measure of the confidence in the assignments. Compounds containing metal–carbon bonds, nitrosyl ligands or metal–ligand multiple bonds have been excluded, and the methods described apply to classic coordination complexes only and not to organometallic compounds. The focus on coordination complexes in part simply reflects the lack of bond-valence data for organometallic compounds, but oxidation-state assignment in organometallic chemistry is also ambiguous; even so fundamental a compound as ferrocene may be considered to contain Fe(0) or Fe(+2).

Two approaches were used in oxidation-state assignment: (i) the bond-valence method and (ii) calculation of ligand charges. The bond-valence method is applicable to any complex whether it is mono- or polynuclear but it depends on the availability of suitable parameters. Ligand charges were derived using three methods: from the minimum of energy versus charge plots, from Pearson's principle of maximum hardness and from automated hydrogen-placement routines. Once the ligand charges are known they can be applied to assign the oxidation state of a metal in a mononuclear complex, but they only yield the total of all the metal oxidation states in polynuclear complexes.

In order to validate and optimize the different methods an initial testing set was generated which contained entries with predefined valences. Suitable entries were extracted from the CSD November 2018 release by scanning for compound names containing a string comprising the name of a transition metal followed by a Roman numeral in parentheses [*e.g.* nickel(II)]. In addition to entries containing one metal, this approach can be applied to multiple metal structures where more than one valence is specified, provided that only one valence is present for each metal name.

The methods described in this article are intended to be applicable to coordination complexes, and so entries containing the organometallic moieties listed in the first paragraph of this section were omitted. 3D co-ordinates were required to be present for all atoms, and disordered structures and structures containing errors were omitted. Entries with missing hydrogen atoms were omitted as well as those where the number of hydrogen atoms present differed from the figure calculated using the CSD structure-editing tools. Where more than one structure is available in a single refcode family, the structure with the lowest *R* factor was selected.

The final test-set contained 54 999 unique metal environments across 47 716 molecular components, from 43 423 entries. This set contained entries from all transition metals, with a minimum number of 52 environments for scandium and a maximum number of 13 259 environments for copper. The test was run twice, first to populate the ligand database with fragment results and then again to enable the charge validation to be incorporated into the confidence scores.

Table 3
Breakdown of assignment results by method against test-set entries.

	Summed component valence assignments			
Method	Correct	Incorrect	% Correct	Not applied
BVS	35 419	3950	89.97	8347
Energy	44 418	3298	93.09	0
Hardness	44 367	3349	92.98	0
Hydrogen placement	41 802	5914	87.6	0
Overall assignment	43 220	4496	90.58	0

The overall success rates of each of the four methods for oxidation-state assignment are summarized in Table 3, where, in order to accommodate both mono- and polynuclear complexes, the entries are based on reproduction of the total metal oxidation state. It should be noted that while application of the ligand-charge methods was achieved successfully for all entries, bond-valence assignments were reliant on the availability of parameters and so have only been applied for 82.50% of components. The ligand-charge methods based on hardness and energy are as effective as the traditional BVS approach and can be applied to all complexes.

The applicability of the BVS method varies significantly across the periodic table with fewer parameters being available for the second- and third-row transition metals than for the first row (Fig. 2). The figures in the first row of Table 3 are thus weighted towards complexes of the 3*d* metals. The BVS method is always needed for assignment of individual metal sites in polynuclear complexes. Therefore, the applicability of the methods described in this article becomes quite patchy for polynuclear complexes of the second- and third-row metals. Of the complexes in Table 3, BVS could not be applied to the metal atoms in 8347 components. Where these components are polynuclear complexes no assignment can be made at all, meaning that no assignment was made for 3113 metal sites in the test set (amounting to 5.7% of the set). This situation should improve as bond-valence parameters are determined for more element bond types in a range of oxidation states, the results obtained here illustrating the importance of research in this area. The oxidation states of the remaining 8347 − 3113 = 5234 mononuclear complexes could all be assigned using the ligand-charge methods.

Some measure of confidence in an oxidation-state assignment can be obtained from (i) the agreement between different methods and (ii) whether any alerts are generated. The success in reproducing author-assigned oxidation states increases substantially over the data presented in Table 3 for the cases where an A or B confidence grade is obtained. Of the 36 080 entries with A assignment, author-assigned oxidation states were reproduced in 99% of cases, with most of the incorrect assignments identifying structural or naming errors in the CSD (Fig. 3).

3.2. Examples

3.2.1. Mononuclear complexes. In the Jahn–Teller distorted 6-coordinate Cu(+2) complex [diaqua-bis(pyrazine-2-carboxylato-*N*,*O*)-copper(II); refcode BEYRAY03 (Wang *et al.*, 2009), Fig. 4(*a*)] BVS parameters are available for Cu—O and Cu—N bonds for oxidation states +1, +2 and +3. As this covers the common copper oxidation states found in the look-up table, BVS is carried out and determines the oxidation state to be +2 with no warnings or errors. The aquo ligand is found to have a charge of zero and the pyridine-carboxylate ligand a charge of −1 by all three ligand-charge methods. When the charges are compared with the fragment charge database no discrepancies are found. No errors or warnings are issued in the ligand-charge-assignment procedure and the total ligand charge of −2 is consistent with the oxidation state assigned by the BVS method. The oxidation state of the Cu is thus assigned to +2 with a confidence score of 17 (A).

A similar process is observed for the nickel complex [bis(2-aminoacetato)-nickel(II) monohydrate, refcode LEPYOV (Wang, 2006), Fig. 4(*b*)], where BVS can be applied to both the common (+2) as well as the less common (+3) oxidation states. A BVS valence of +2 is determined and the total ligand-charge calculations are consistent with this for all methods. This assignment is awarded the highest confidence score: 17 (A) with no discrepancies between these and previous ligand-specific assignments.

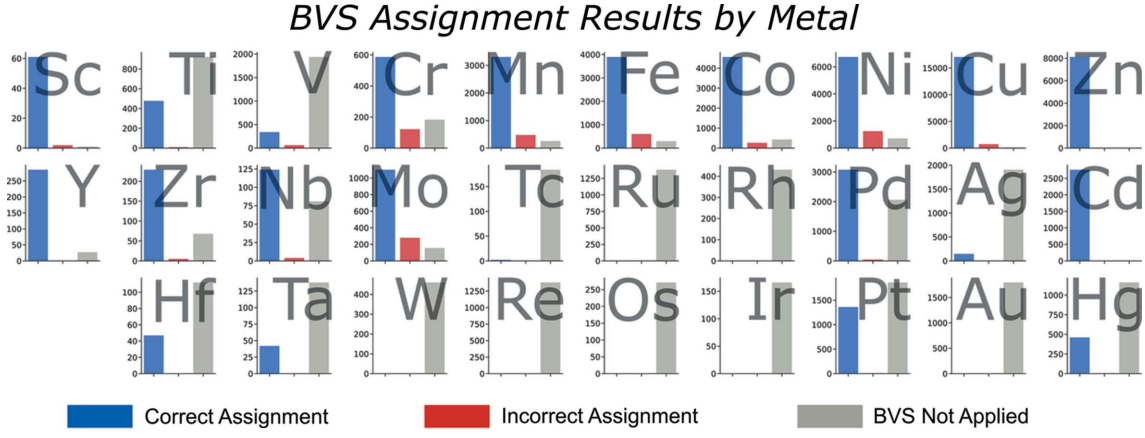


Figure 2
Application of the BVS method by metal to test-set entries. Bars show relative success/failure and applicability of BVS for each of the first-, second- and third-row transition-metal atoms. BVS is not applicable when metal–ligand specific parameters are lacking the common oxidation states of that metal.

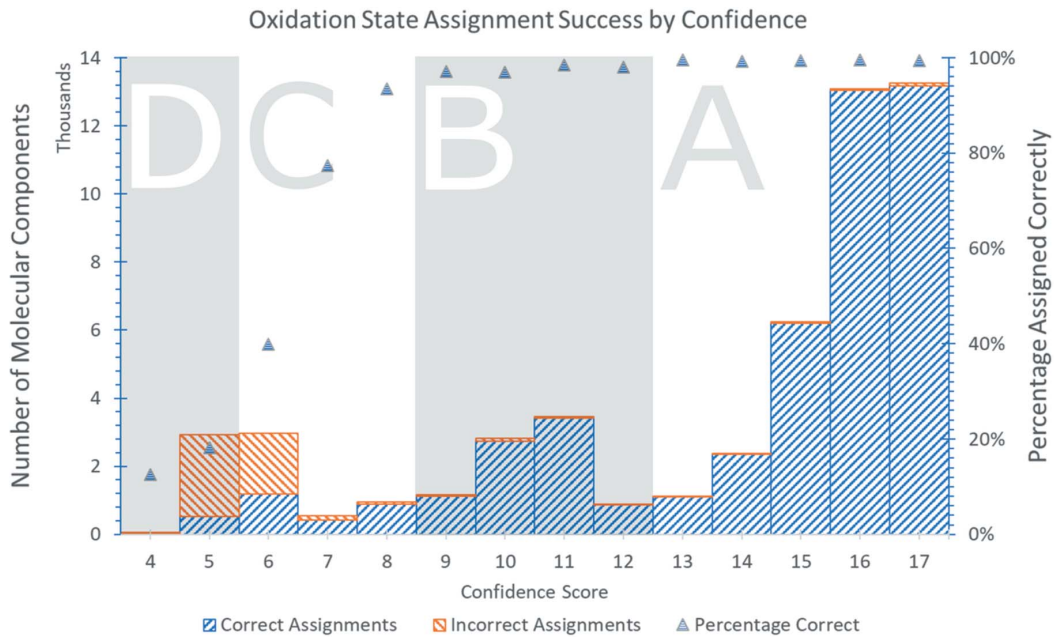


Figure 3
The success rate of oxidation-state assignment grouped by confidence-score bands. The raw data for this figure are available in the supporting information (Table S4).

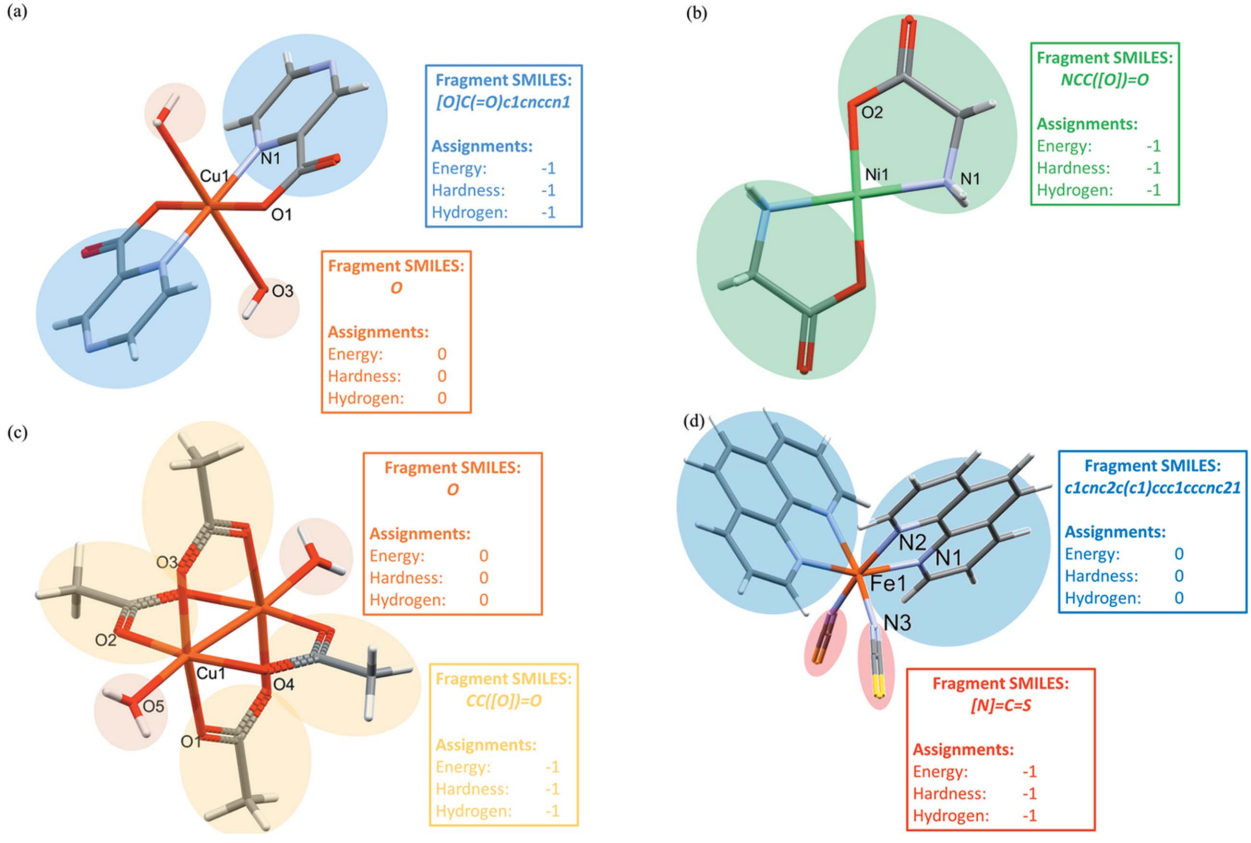


Figure 4
Oxidation-state assignment examples. Boxes illustrate individual ligand fragments with corresponding SMILES notation and assignment-method results demonstrated. Note only asymmetric unit metal–ligand atoms are labelled for clarity. (*a*) Refcode BEYRAY03, a copper(+2) structure with Jahn–Teller distortion; (*b*) refcode LEPYOV, a planar nickel(+2) complex; (*c*) refcode CUAQAC01, copper(+2) acetate with Cu–Cu bond as depicted in the CSD; and (*d*) refcode KEKVIF, a family of iron(+2) compounds that exhibit SCO behaviour.

3.2.2. Metal–metal bonds. The identification of metal–ligand bonds has been based on inbuilt CSD functions for defining bonded atoms. The algorithm also generates bonds for short metal–metal distances. Metal–metal bonding is a widely studied area of organometallic chemistry, but entries containing metal–carbon or metal–ligand multiple bonds were excluded from this study and so many entries with metal–metal bonds would have been omitted on this basis. Nevertheless, the CSD bonding criteria generate metal–metal bonds in coordination complexes such as the copper acetate dimer [tetrakis(μ^2-acetato)-di(aqua)-di[copper(II)], refcode CUAQAC01 (Mahmoudkhani & Langer, 1998), Fig. 4(*c*)] in which the Cu···Cu distance is 2.619 Å. While metal–metal bond formation does not affect the ligand-charge procedures, the BVS method would fail because metal–metal bonds are not present in the bond-valence parameter database used in this study.

Short metal–metal distances in coordination compounds are usually the result of the geometric demands of bridging ligands rather than genuine metal–metal bonding. We have simply omitted metal–metal bonds from the BVS calculations. The BVS calculation can then proceed as usual, yielding in the case of refcode CUAQAC01 a value of +2 for the copper oxidation state. The assignment is supported by each of the ligand-charge methods. The confidence score is 17(A).

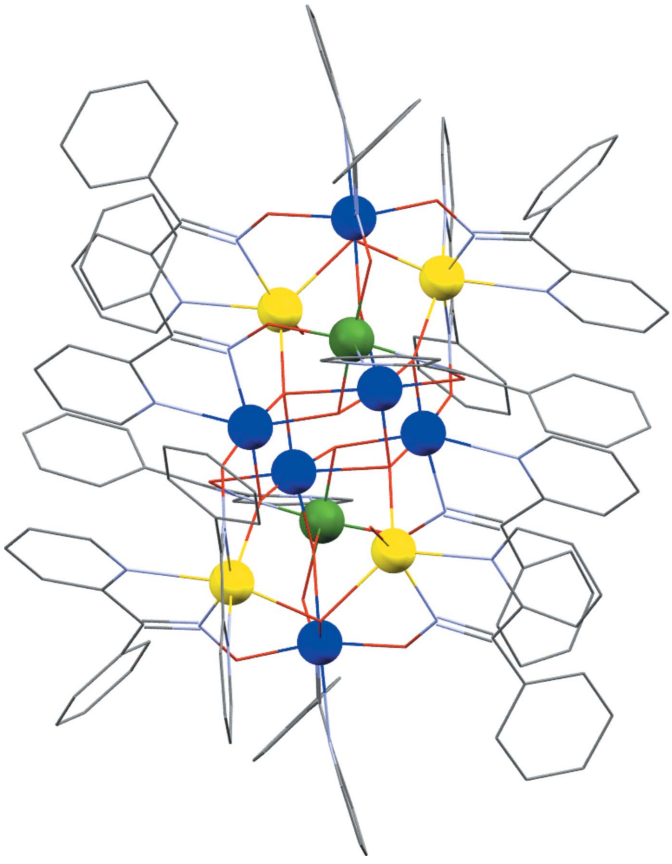


Figure 5
Mn12 structure (refcode ZAVMEQ) containing six unique metal centres. Mn(+2) is in yellow, Mn(+3) is in blue and Mn(+4) is in green. H atoms are omitted for clarity.

3.2.3. Spin cross-over complexes. The adoption of a high- or low-spin configuration affects metal–ligand bond distances and can influence oxidation-state assignment via the BVS method. The crystal structures of many materials of interest in terms of spin cross-over (SCO) behaviour have been determined in multiple spin states and occur in the CSD as refcode families where entries have the same six letter code but differ in the last two digits. While this test has focused on a single entry from each refcode family (see Section 3.1), to understand the role of SCO on oxidation-state assignment the process has been applied to a family of structures with both spin states present.

Complexes of Fe(+2) with nitrogen ligands have been widely investigated, and in this case both high-spin and low-spin bond-valence parameters are available. For example, the refcode family for the iron complex [*cis*-bis(isothiocyanato)-bis(1,10-phenanthroline-*N*,*N*′)-iron(II), refcode KEKVIF (Gallois *et al.*, 1990), Fig. 4(*d*)] contains nine entries with atomic coordinates. The ligand-charge assignment methods produce the same result in each case, determining the thiocyanate and phen ligands to have charges of −1 and 0, yielding a metal oxidation state of +2. While the BVS method assigns an incorrect oxidation state of +3 for the low-spin entries the unrounded values are all above 3.5, which generates a warning message. As a result of the discrepancies between BVS and ligand-charge assignment methods, along with the warning in the BVS assignment, the low-spin complexes (such as refcode KEKVIF02; Granier *et al.*, 1993) have a very low confidence score of 5(D). This situation occurs commonly with SCO families, and as such, SCO families are identifiable by large differences in confidence between entries.

For future assignments an extra parameter has been added to the confidence score which warns of potential refcode family issues. Where an entry is part of a refcode family, metal-bond distances are checked across the whole family when assigning oxidation states. If metal–ligand bond distances vary by more than 0.1 Å within the same refcode family, the confidence score is reduced by four and a warning is issued.

3.2.4. Mixed-valence polynuclear complexes. The assignment of oxidation states in polynuclear complexes is based on a combination of BVS and ligand-charge methods. The Mn complex in Fig. 5 (refcode ZAVMEQ; Alexandropoulos *et al.*, 2012) contains six unique metal centres with oxidation states between +2 and +4. The BVS method matched the literature values for all six metal centres, with no warnings or errors. The summation of ligand charges was −34 for all methods, which is consistent with the BVS total +34. In the ligand-charge calculation for the pyridinyl-methanimine ligand there is a small (<1 eV) difference in the hardness for charges of −1 and −3, which generates a warning. A warning is also produced for shallow energy curves for some fragments. The confidence score is 15(A).

The CSD compound-naming conventions mean that a complex containing a single metallic element in multiple oxidation states would not have been part of the test dataset and instead the procedures were validated manually by comparing assignments with those given in the corresponding

publications. Approximately 50 complexes were examined over the course of this work and four errors were identified. In each case, the total of BVS assigned oxidation states did not match the ligand-charge assignment methods and therefore all received a low confidence score of 5/6.

3.2.5. Complexes with open-shell ligands. The closed-shell restriction applied during the ligand-charge calculations means that ligands with odd numbers of electrons are incorrectly treated. Complexes containing radical ligands were identified in the test dataset by comparing the author-assigned oxidation states with the allowed closed-shell charges for all ligands. A radical ligand is present if the named oxidation state does not match possible open-shell charges. This was found to be the case for fewer than 2% of structures. Moreover, most of the instances involved a small set of common radicals. The SMILES formulae of identified radicals have been added to the ligand SQLite database. For example, nitroxide radicals are identified from SMILES string segments CN([O])C, cN([O])c, cN([O])C and CN([O])c. It is additionally possible to add complete ligand-specific SMILES manually for individual radical ligands where needed.

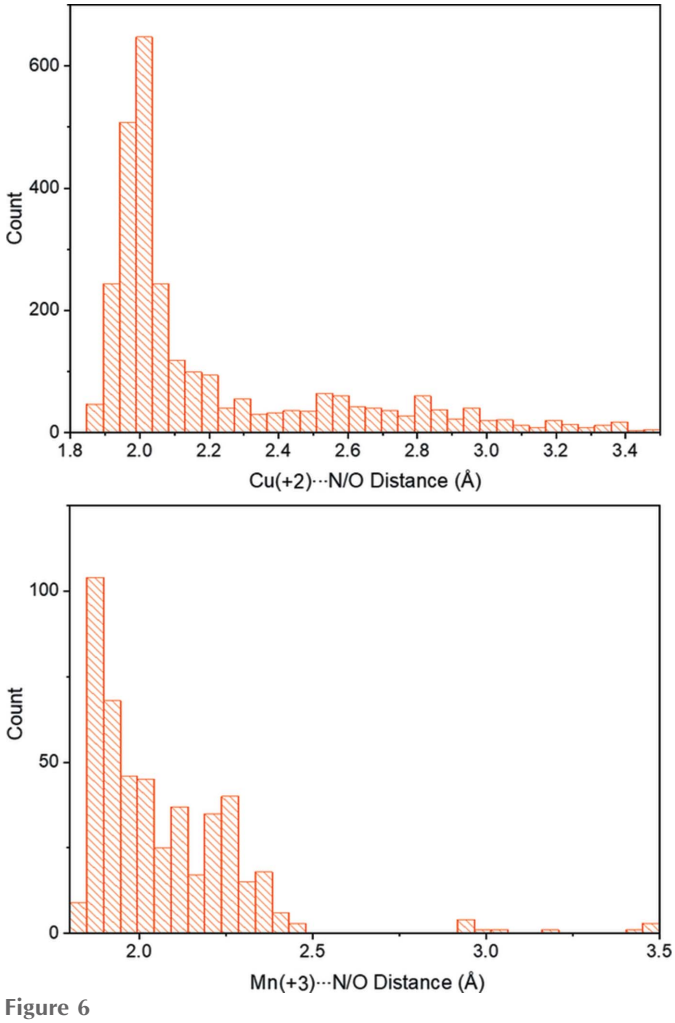


Figure 6
Distributions of Cu(+2)–N/O (left) and Mn(+3)–N/O (right) interatomic distances from the CSD.

There is a tendency, in the case of open-shell ligands, for the energy and hydrogen-placement charge assignments to suggest values of −1, while hardness often suggests a charge of +1. BVS assignments are usually correct. This disagreement results in low confidence scores (typically D) where radical containing complexes are encountered for the first time. For example, in the dinuclear 1,2,3,5-diselenadiazolylnickel(+2) complex BARXID (Wu *et al.*, 2012) BVS correctly assigns the oxidation state as +2 for both metal sites. The radical is given a charge of −1 according to energy and hydrogen placement, and +1 according to hardness. Overall the confidence score is 6 (C): 6 for the BVS method and zeros for all the ligand-charge criteria.

3.2.6. Demonstration of oxidation-state specific data: the Jahn–Teller effect in Cu complexes. With atom-specific valences now available, it is possible to limit some common geometric searches to specific oxidation states. For example, the availability of atom-specific oxidation-state data enables rapid collation of a list of Cu–ligand bond distances in Cu(+2) sites. Fig. 6 shows a histogram of the distances obtained from copper sites with at least six short (< 3.5 Å) Cu···N/O contacts where the oxidation-state assignment has a confidence of A or B. The expected bimodal distribution between 1.8 and 2.8 Å shows elongation of metal–ligand bond lengths for axial ligands. The plot enables an upper limit of about 3.0 Å to be placed on a Jahn–Teller axis in these complexes. A similar pattern is observed for Mn(+3) structures, with a bimodal distribution suggesting the same Jahn–Teller distortion out to 2.5 Å.

4. Conclusions

The aim of the methods described in this article is to automate assignments of oxidation states to metal sites in mononuclear and polynuclear coordination complexes in the CSD. Each assignment is given a confidence score. Assignments with scores of A or B appear to be reliable, yielding the correct assignment for 99% of cases during testing. Assignments with scores of C and D often represent special electronic or bonding situations, such as non-innocence associated with redox-active ligands, spin-state ambiguity or open-shell ligands. These cases still require manual checking. Experience over the course of this project suggests that the ultimate aim of completely automated oxidation-state assignment without any manual intervention at all would be difficult or impossible to meet when based only on structural data.

The methods developed and investigated here will be implemented as part of the curation process of the CSD by expert scientific editors at the CCDC. In this manner, oxidation states where there is reliability in the assignment and/or clear pre-assignment by the author will be transferred straightforwardly into the curated CSD entry. The focused attention from scientific editors can then be applied to the structures where the assignment is less reliable or indicates some interesting or unusual chemistry.

A project is also currently underway to evolve the format of the CSD and this will allow automated transfer of oxidation

states from the compound name in the entry (current) to be an atomic property on individual metal sites (future). The approaches described here will certainly help in that translation as well.

The availability of site-specific oxidation states as searchable criteria in the CSD should enable more targeted applications of the database in transition-metal and materials chemistry. It should be possible, for example, to investigate how a metal and its oxidation state determine the deformability or structural flexibility of coordination; such information could be helpful in the design of metal-templating reactions. Complexes with sites exhibiting unusual geometries might be susceptible to modification by high pressure or irradiation. The combination of oxidation-specific searching with motif-searching tools such as the Crystal Packing Feature component in *Mercury* (Macrae *et al.*, 2008; Childs *et al.*, 2009) may find uses in research aiming to establish the relationship between topology and magnetic properties. Finally, the SQLite ligand database could be extended to include a variety of properties such as conformational flexibility, pK_a, number of donor sites, *etc*. that may be helpful in ligand design.

5. Available stand-alone software

Although the methods described above are designed to work with curated entries in the Cambridge Structural Database, a stand-alone script, named *MRMOX*, can be downloaded from the link http://www.crystal.chem.ed.ac.uk/software/mrmox. The script works through the *Mercury* CSD Python API menu to assign oxidation states with input from users' own cifs. A short set of installation and usage instructions is available in a READ_ME file in the download. The program will only work under *Windows* with a licensed installation of the CSD, including *Mercury* and the CSD Python API.

Acknowledgements

We thank Professor I. D. Brown (McMaster University, Hamilton, Canada) and Professor J. J. P. Stewart (Stewart Computational Chemistry, Colorado Springs, USA) for very helpful discussions and advice.

Funding information

We thank the Cambridge Crystallographic Data Centre and the Engineering and Physical Sciences Research Council for studentship funding to MGR.

References

Akbarzadeh Torbati, N., Rezvani, A. R., Safari, N., Saravani, H. & Amani, V. (2010). *Acta Cryst.* E**66**, m1284.
Alexandropoulos, D. I., Manos, M. J., Papatriantafyllopoulou, C., Mukherjee, S., Tasiopoulos, A. J., Perlepes, S. P., Christou, G. & Stamatatos, T. C. (2012). *Dalton Trans.* **41**, 4744–4747.
Brown, I. D. (2016*a*). *The Chemical Bond in Inorganic Chemistry: The Bond Valence Model.* 2nd ed. Oxford University Press.
Brown, I. D. (2016*b*). *Bond valence parameters*, https://www.iucr.org/resources/data/datasets/bond-valence-parameters.
Chérif, I., Abdelhak, J., Zid, M. F. & Driss, A. (2013). *Acta Cryst.* E**69**, m667–m668.
Childs, S. L., Wood, P. A., Rodríguez-Hornedo, N., Reddy, L. S. & Hardcastle, K. I. (2009). *Cryst. Growth Des.* **9**, 1869–1888.
Das, A., Basuli, F., Peng, S.-M. & Bhattacharya, S. (2002). *Inorg. Chem.* **41**, 440–443.
Davies, D. W., Butler, K. T., Isayev, O. & Walsh, A. (2018). *Faraday Discuss.* **211**, 553–568.
Gallois, B., Real, J. A., Hauw, C. & Zarembowitch, J. (1990). *Inorg. Chem.* **29**, 1152–1158
Granier, T., Gallois, B., Gaultier, J., Real, J. A. & Zarembowitch, J. (1993). *Inorg. Chem.* **32**, 5305–5312
Groom, C. R., Bruno, I. J., Lightfoot, M. P. & Ward, S. C. (2016). *Acta Cryst.* B**72**, 171–179.
Hipp, D. R. (2019). *SQLite*, Version 3.29.0. Wyrick & Company Inc., Charlotte, NC, USA.
Holler, S., Tüchler, M., Belaj, F., Veiros, L. F., Kirchner, K. & Mösch-Zanetti, N. C. (2016). *Inorg. Chem.* **55**, 4980–4991.
Housecroft, C. E. & Sharpe, A. G. (2008). *Inorganic Chemistry*, 3rd ed. Harlow, UK: Pearson Prentice Hall.
Macrae, C. F., Bruno, I. J., Chisholm, J. A., Edgington, P. R., McCabe, P., Pidcock, E., Rodriguez-Monge, L., Taylor, R., van de Streek, J. & Wood, P. A. (2008). *J. Appl. Cryst.* **41**, 466–470.
Mahmoudkhani, A. H. & Langer, V. (1998). *CSD Communication*.
Mezei, G. & Raptis, R. G. (2004). *Inorg. Chim. Acta*, **357**, 3279–3288.
Pearson, R. G. (1993). *Acc. Chem. Res.* **26**, 250–255.
Rotondo, E., Cusmano Priolo, F., Bombieri, G. & Bruno, G. (1984). *Acta Cryst.* C**40**, 960–962.
Shields, G. P., Raithby, P. R., Allen, F. H. & Motherwell, W. D. S. (2000). *Acta Cryst.* B**56**, 455–465.
Stewart, J. J. P. (2013). *J. Mol. Model.* **19**, 1–32.
Stewart, J. J. P. (2016). *MOPAC2016*, http://openmopac.net.
Torzilli, M. A., Colquhoun, S., Doucet, D. & Beer, R. H. (2002). *Polyhedron*, **21**, 697–704.
Wang, G.-H., He, R.-L., Meng, F.-J., Hu, N.-H. & Xu, J.-W. (2009). *Acta Cryst.* E**65**, m1511.
Wang, Z.-L. (2006). *Acta Cryst.* E**62**, m2546–m2548.
Wu, J., MacDonald, D. J., Clérac, R., Jeon, I. -R., Jennings, M., Lough, A. J., Britten, J., Robertson, C., Dube, P. A. & Preuss, K. E. (2012). *Inorg. Chem.* **51**, 3827–3839.