Using collective knowledge to assign oxidation states

Kevin Maik Jablonka, Daniele Ongari, Seyed Mohamad Moosavi, and Berend Smit*

Laboratory of Molecular Simulation (LSMO), Institut des Sciences et Ingenierie Chimiques (ISIC), École Polytechnique Fédérale de Lausanne (EPFL), Sion, VS, Switzerland.

Knowledge of the oxidation state of a metal centre in a material is essential to understand its properties. Chemists have developed several theories to predict the oxidation state on the basis of the chemical formula. These methods are quite successful for simple compounds but often fail to describe the oxidation states of more complex systems, such as metal-organic frameworks. In this work, we present a data-driven approach to automatically assign oxidation states, using a machine learning algorithm trained on the assignments by chemists encoded in the chemical names in the Cambridge Crystallographic Database. Our approach only considers the immediate local chemical environment around a metal centre and, in this way, is robust to most of the experimental uncertainties in these structures (like incorrect protonation or unbound solvents). We find such excellent accuracy ($>$ 98 %) in our predictions that we can use our method to identify a large number of incorrect assignments in the database. The predictions of our model follow chemical intuition, without explicitly having taught the model those heuristics. This work nicely illustrates how powerful the collective knowledge of chemists actually is. Machine learning can harvest this knowledge and convert it into a useful tool for chemists.

Main

Oxidation states are a concept every chemist learns, at the latest, in their first days as undergraduates. Their history goes back to the early days of chemistry when Lavoisier coined the word oxidation and Wöhler the expression "oxydationsstufe" (old German spelling for the term oxidation number)[1;2]. Oxidation states are central to balance redox reactions[3], for chemical nomenclature[4], and above all to help chemists to systematise and reason about (redox) reactivity as well as spectroscopic properties[5–7]. The concept of oxidation states plays such an important role in the fundamentals of chemistry that some have argued that the oxidation numbers should be represented as the third dimension of the periodic table[8].

Every chemist also experienced that assigning oxidation states is not trivial. The International Union of Pure and Applied Chemistry (IUPAC) defines oxidation states as "...the charge of this atom after ionic approximation of its heteronuclear bonds ..."[9;10]. This definition is, however, too generic and cannot be readily translated into a recipe to determine the oxidation state of any given compound. Therefore, in practice, chemists fall back to formal electron counting rules. For molecules, this approach gives satisfactory results for most cases. For crystalline materials, however, these electron counting rules often fail as they are based on bonds and bond orders, which are ill-defined for crystalline materials[11].

Therefore, for crystalline materials the oxidation state is often estimated using the bond valence sum method[12]. This method, which dates back to Linus Pauling[13], approximates all bonds as fully ionic, and the oxidation state is estimated by summing up all bond valence sums, which are calculated based on a parametrization of an exponential function of metal-ligand bond lengths. There is an ongoing effort in tuning the bond valence sum method to being able to automatically evaluate the entries in the Cambridge Structure Database (CSD)[14;15], which is the largest collection of metal-organic crystals.

The bond valence sum method is, however, far from ideal as it has many ambiguities. First, one needs to assign bonds between the atoms, for which there is no unique procedure for crystalline materials[11;16]. Second, a large number of different parameter sets exist which are derived for different classes of materials, e.g., metal oxides or metal-organic complexes[12]. There is little insight, if any, in the transferability of these parameters to novel classes of materials, e.g., metal-organic frameworks (MOF). Yet, these parameters are often mixed to cover chemical space[12], and different groups may use different parameters for the same set of materials.

Finally, the functional form for the bond valence method might sometimes be too rigid as it is based solely on bond lengths. This can cause fundamental problems for a number of systems, e.g., in the case of non-innocent ligands for which the electron donation count depends on the coordination geometry, in sterically constrained systems for which bond lengths can be outside the expected range, or for less well-defined coordination polyhedra[17–19].

In this context, it is important to note that quantum chemical calculations are of limited use. From a fundamental point of view, one could argue that a state-of-the-art quantum chemical calculations would give us the total energy for the different oxidation states, and hence it would be straightforward to determine the oxidation state that gives the lowest energy. Unfortunately, for most MOFs the unit cell is so large that one has to use density functional theory (DFT), which tends to favor compounds with lower d orbital occupancy and leads to non-integer oxidation states for multivalent compounds, like magnetite (Fe_3O_4), due to the self-interaction error (in the generalized gradient approximation (GGA))[20;21]. Other computational techniques have been developed that are based on charge-partitioning schemes, but as the charge on an embedded atom is not well defined, and subject to charge-transfer interactions with the ligands[8;22], also these methods are not able to remove the ambiguity in the assignment of oxidation states. Because of these difficulties, most, if not all, quantum calculations on MOFs and other materials with unit cells that have a large number of atoms, in fact, require a "guess" of the oxidation state as input, rather than giving us insight in the actual oxidation state.

In summary, there is a need for a new approach towards the assignment of integer oxidation states that is able to capture the intricacies of chemistry—to provide starting points for DFT calculations and to support chemical reasoning.

In this work, we propose to use the collective knowledge of chemists to assign oxidation states, replacing the rule-based deductive approach of formal counting rules and the bond valence method with a fully inductive one. Our approach harvests the collective knowledge of thousands of chemists to create a consensus assignment of oxidation states, which to our knowledge has not been explored to provide a simple solution to this important practical question.

In our approach, we parsed the chemical names in the CSD for oxidation states of metal centres, numerically encoded the local chemical environment, and trained an ensemble of machine learning (ML) models to classify the oxidation state. We chose to focus on MOFs as their experimental structures are archetypal examples for many of the reasons deductive techniques might fail to assign oxidation states: Unbound solvent molecules are present in many experimental MOF structures, and sometimes the structures also contain charge compensating counterions. Moreover, our model is challenged by problems like missing or incorrect protonation as well as atomic disorders. Even if our main focus in this work is on predicting the oxidation state of metal centres in MOFs, we also demonstrate that our model that was trained only on MOFs is able to transfer to other types of chemistry.

Results and discussion

To create our data set of oxidation states for metal centres in MOFs, we leveraged the fact that the chemical names of nearly half of all entries in the CSD[23] contain oxidation

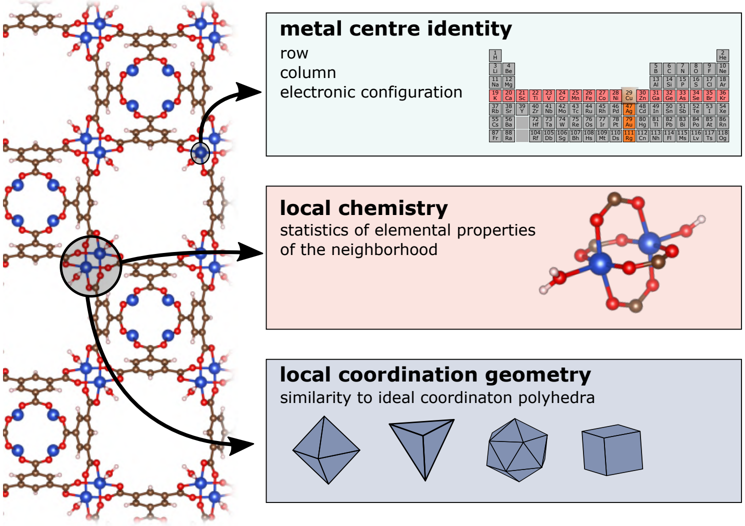
metal centre identity
row
column
electronic configuration
local chemistry
statistics of elemental properties
of the neighborhood
local coordination geometry
similarity to ideal coordinaton polyhedra

Fig. 1 | Schematic representation of the featurization approach. The local chemistry features are based on statistics of elemental properties. The local geometry is captured by measuring the similarity of the actual coordination environment to ideal coordination environments.

states in parentheses following the metal names (as it is recommended in the guidelines for inorganic nomenclature, the IUPAC red book)[4]. This assignment can be based on different arguments: chemical intuition, founded on knowledge of the chemical literature and experience with similar reactions and compounds, some computing protocol (e.g., the bond valence method), or spectroscopic evidence. Even if these oxidation states are not assigned with a unique and well-defined protocol, several chemists (at least the authors and the editor at the CSD) consider this assignment to be correct[23;24]. The central assumption in this work is that individual assignments might be wrong but if enough chemists work on similar systems the collective knowledge will be right.

Encoding local environments and machine learning

In this work, we use a ML model to capture the collective knowledge of chemists on the oxidation state. To be able to train ML models, one has to encode the local environment as a vector of numerical descriptors ("features"). This is commonly known as featurization, and the success of any ML model crucially depends on selecting features that are able to describe the problem at hand[25]—ideally in a physically meaningful way[26]. We based our featurization approach on the locality approximation in which we consider only the immediate local environment around a metal centre in a structure (cf. Fig. 1 for an illustration). This is also reflected in Pauling's principle of local charge neutrality[27], as well as the nearsightedness principle of electronic matter, which describes that the density change caused by a potential change far an away is small[28]. In addition to being physically meaningful, this approximation allows us to create a large training set that enables powerful similarity-based reasoning. This realization reflects Pauli's parsimony principle which states that the number of unique local environments is limited. Using the locality approximation, we can also consider structures with unbound solvent molecules and missing, or incorrect, protonation as those solvent molecules or missing protons are typically outside the local environment of a metal centre.

Our feature vector combines the three aspects chemists have identified as key to the oxidation state: the metal type, the chemical environment, and the geometry of the coordination environment (see Fig. 1). We used the first two values of the feature vector to identify the position of the metal in the periodic table, i.e., its row and group number. The column encodes the well-known principle that elements in the same group share similar chemistry, and the addition of the row makes the encoding of the metal position in the periodic table unique. We further added the number of electrons in the different shells as additional features for the metal centre.

The next elements of our feature vector recognise that there is a deep relationship between coordination geometry and the electronic configuration. Prime examples for this relationship are the ligand field splittings for different coordination environments and the Jahn-Teller distortion for degenerate electronic configurations. To encode these effects numerically, we use order parameters, which measure the similarity of the coordination environment to a collection of ideal coordination environments (e.g., octahedral, tetrahedral, bent linear). In this way, we capture heuristics like "square-planar Pt is usually d^8" which experienced chemists can rely on but which are difficult to comprehensively encode in a deductive approach. Importantly, our featurization does not explicitly depend on bond lengths, which makes it more robust and interoperable—e.g., we can use the same model on DFT optimized and experimental structures.

The key insight on which formal counting rules are built upon is that different ligands are thought to donate a different number of electrons. We attempted to encode this more flexibly by calculating statistics, like the electronegativity differences, of elemental properties between the metal centre and its geometrical nearest neighbors[29].

The matrix describing the immediate local environments was then used as an input for a voting classifier which arrives at its final prediction by averaging the predictions (probability of oxidation states) of four base models, each based on a different approach (decision trees, nearest neighbours, and linear functions in feature space). The use of this voting makes our predictions more robust and provides us with an uncertainty estimate[30]. This approach is similar to the way in which we use the collective knowledge of chemists at the level of the training data to arrive at a data-driven definition for the oxidation state—not all chemists use the same method to assign the oxidation state but taken together the collective assignment for a particular chemical environment can be robust.

After calculating the feature vector for each metal site, we split the data into disjoint sets for training and testing (see Computational Methods section for more details). In addition to that, we also use structures with strong spectroscopic evidence for the oxidation state assignment as separate test cases.

Performance assessment

To assess the accuracy of our method, we focused on copper, for which we can compare our results with an optimized and validated bond valence method[14]. In addition, for

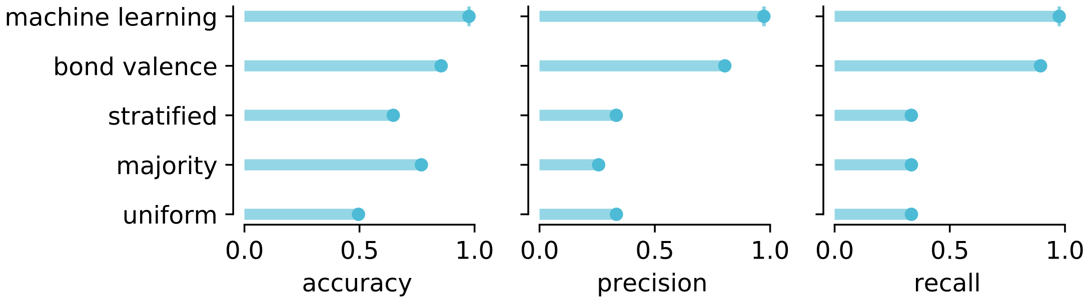

Fig. 2 | Performance metrics. Accuracy, precision, and recall for the assignment of oxidation states for Cu in the MOF subset of the CSD using different classifiers that draw from a uniform distribution (i.e., same probability for both oxidation states), only the majority class (i.e., all structures are assigned II), the training set distribution (stratified, i.e., assigning Cu(II) with a ca. 75 % chance), the bond valence sum method as well as our ML model.

copper both oxidation states I and II are well represented in the MOF subset of the CSD (Cu(I): 24.2 %, Cu(II): 75.8 %). To determine the performance of our ML method and the optimized bond valence method, we calculate the accuracy of our predictions as well as measures that are sensitive to the number of false positives (precision) and false negatives (recall). Due to the imbalanced distributions of Cu(I) and Cu(II) we already have a 75 % chance of success by assuming all oxidation states to be II (majority vote in Fig. 2). Similarly, we can perform a random or stratified random assignment of the oxidation state. These models are important as a baseline for the performance metrics. Fig. 2 clearly shows that our model outperforms the baselines and the bond valence method in all metrics.

It is interesting to use our ML results to investigate why the bond valence method fails for some structures. For this, we projected our feature space onto two dimensions using principal component analysis (PCA, cf. Fig. 3). In these principle components, the two most relevant feature values are the extent to which the copper is trigonal co-planar with coordination number three, and the extent to which copper is square co-planar with coordination number four. The black stars are structures for which the

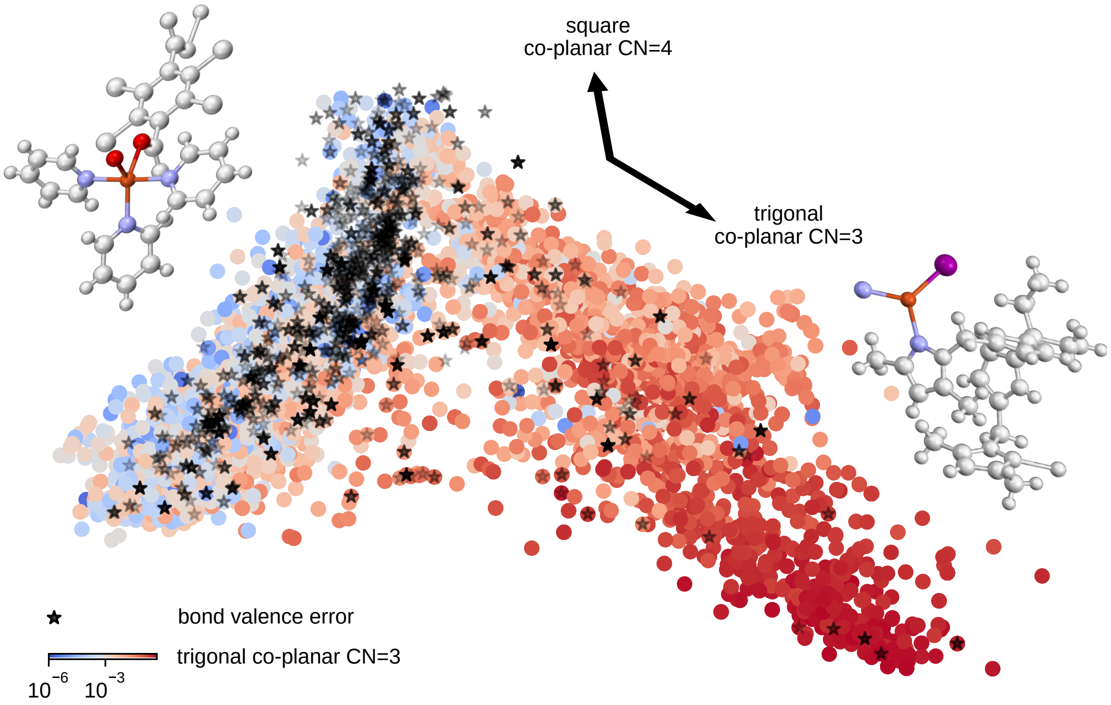

Fig. 3 | First two principal components for Cu sites. Projection of the feature feature space onto the two first principal components (linear combinations of features that capture most of the variance in the data). The colour coding shows the value of the order parameter of the trigonal planar coordination (logarithmically scaled colour map). Black stars mark metal sites of structures for which the bond valence method predicted the wrong oxidation state. The two arrows denote the directions of the two original features that have the highest contribution in the first and second principle component, respectively. We also show two structures that are at the extremes of the first principal component and for which the bond valence sum method is wrong and correct, respectively.

bond valence method predicts the oxidation state incorrectly. We can see that these incorrect assignments cluster for copper with high coordination numbers. In our model, we see that for these structures the geometric features are of higher importance, and exactly these geometric features can not be described in the distance-based bond valence approach.

By design, our method is directly applicable to all metals. To obtain a more detailed measure of the success of our predictions, we used a test set of 42,463 metal sites that

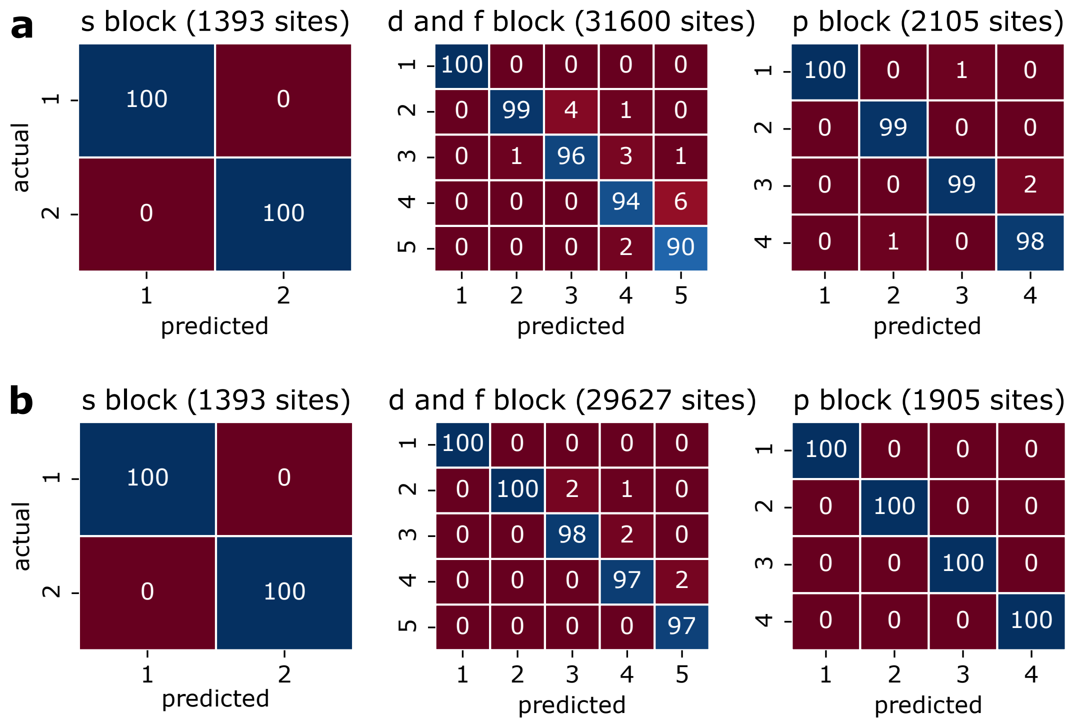

Fig. 4 | Predictive performance across the periodic table. **a**, Confusion matrices for all predictions, independent of the uncertainty of the model. **b**, Confusion matrices only for predictions for which all base estimators agree (39,943 sites, s block: 100 %, d and f block: 94 %, p block: 90 %). Confusion matrices calculated for predictions on a holdout test set of 42,463 metal sites. Annotations in the confusion matrix in percent, which is also used for the colour coding.

were not used in the training set to compute the confusion matrices for different parts of the periodic table (cf. Fig. 4**a**). For the s block (e.g., Li, Na, Ca) all oxidation states were correctly assigned. Even for the more challenging d block (e.g., Fe, Cu), p block (e.g., Al, Pb, Bi), and f block (e.g., Ce, Eu, Ho) we obtained success rates of at least 90 %. These results translate in commonly used classification metrics such as (balanced) accuracy that exceed 98 % (see Supplementary Information).

One additional advantage of ensemble ML models is that they can provide information on how reliable a prediction is. Models based on different hypothesis spaces tend to disagree when used outside the domain of applicability (i.e., when they extrapolate) and agree when the queried case is well represented in the training data. For our model, we find a mean difference between the number of disagreeing base estimators of 0.89, which indicates that usually one base estimator will disagree in the case of a wrong prediction. If we use this to eliminate predictions in which our model is uncertain, we find that the overall prediction accuracy increases significantly. We now also get near-perfect predictions for the p, low valence d, and f block metals (see Fig. 4**b**).

It is instructive to investigate those cases in which we make a prediction with a high confidence, yet make a wrong assignment. These structures (ca. 300) were flagged and we retrieved the article to manually inspect the oxidation state. Out of these, in 70 cases we observed that the assignment in the CSD did not match the one in the original paper (see Supplementary Information), often caused simply by the exchange of IV to VI or I with to II. In the rest of the articles, the oxidation state was either based on experiments or guessed by the authors. For several of them we question the assignment of the oxidation state and, of course, we also have cases in which our method incorrectly assigns the oxidation state. All these cases are listed in the Supplementary Information. The fact that the majority of the cases with discrepancies are erroneous assignments in the CSD suggests that it would be advantageous to use our method as a diagnostics; if we make a high confidence prediction that differs, a more detailed investigation into the oxidation state would be advisable.

To further confirm the accuracy of our predictions, we identified a number of structures for which the oxidation state assignment is supported by strong spectroscopic evidence. Also here, the model showed a good performance by predicting the correct results in all but one of over 50 cases, including mixed-valence cases. In the Supplementary Information these cases are listed. The one structure for which our method failed is a MOF for which it is known that there are missing linker defects[31;32], but the unit cell that was provided gave an averaged equivalent environment of all metals and hence our model gave one oxidation state.

Case studies

It is interesting to look at the assignments of a few case studies in detail. Of particular interest are MOFs with mixed-valence and the case of flexible MOFs for which there is considerable discussion in the literature about the oxidation state.

The importance of geometrical features in assigning the oxidation state is evident for the case of the mixed-valence MOF Cu(I/II)-BTC. Mixed-valence MOFs have been excluded from our training set as the CSD does not systematically indicate which oxidation state corresponds to which metal. As our features are local, we can use the model to determine the oxidation state for each metal site in these mixed-valence MOFs. Since our program does not consider the symmetry, we determine the oxidation state for each of of the 12 metal sites in the Cu(I/II)-BTC unit cell separately (cf. Fig. 5). In agreement with the experimental data[33] we assign the four coppers in the paddlewheel to be +II while the eight coppers in the macrocyclus are assigned to be +I. In Fig. 5, we also illustrate the relative importance of the different features that determine the assignment. The assignment is mostly based on the local coordination geometry (blue ring), where for the paddlewheel the square pyramidal (sq. pyr.) and for the macrocyclus the linear order parameter is the most important feature. In Fig. 5**e** and **f** we give the top five features that determine the oxidation state. In these figures, each dot corresponds to one of the 12 metal sites. If the structure would be perfectly symmetric, there would be only two dots. Our order parameters for the coordination environment reflect that the Cu in the paddlewheel is considerably square pyramidal (high values for sq. pyr.) but not linear, while the opposite is true for the Cu in the macrocyclus. This nicely illustrates how our model captures the chemical intuition that a square pyramidal coordination environment is always associated with Cu(II) whereas a linear coordination environment is associated with Cu(I).

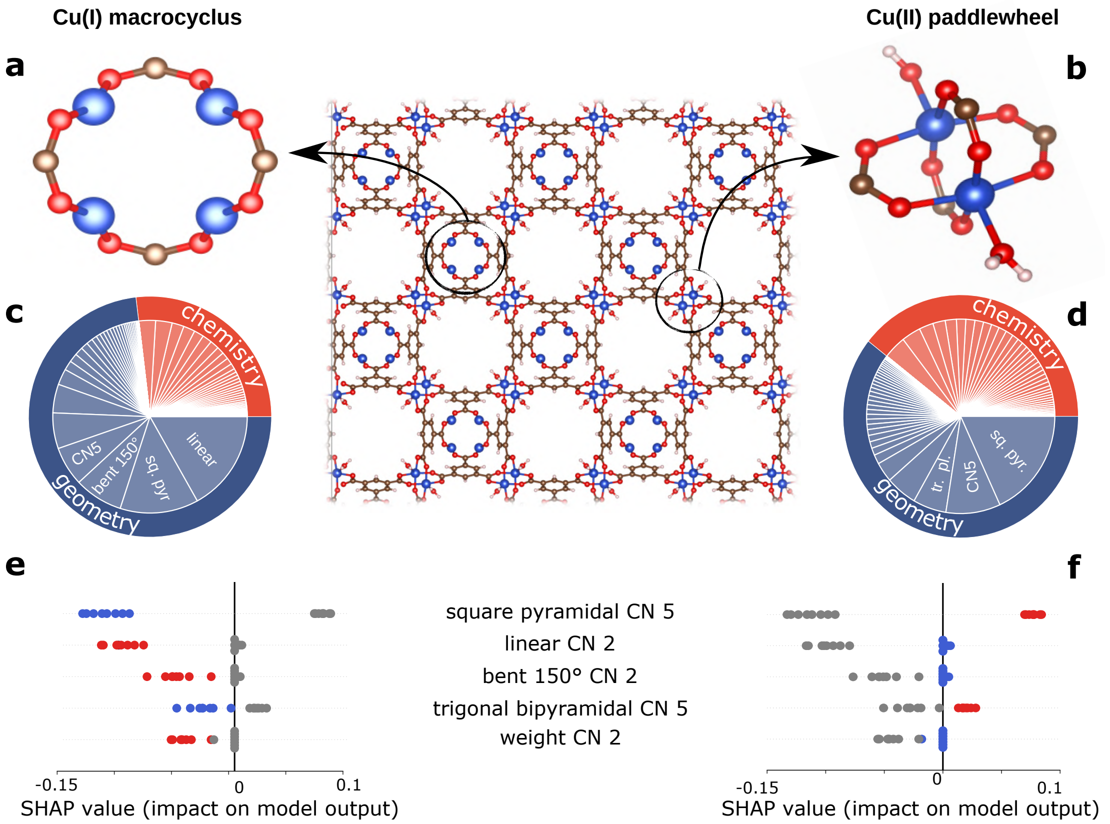

Fig. 5 | Predictions of the oxidation states in mixed-valence MOF Cu(I/II)-BTC. **a**, Cu(I) macrocyclus. **b**, Cu(II) paddlewheel. **c** and **d**, Global feature importance for the Cu(I) and Cu(II) sites, respectively. In both cases, the geometrical features (CN: coordination number) are of highest importance for the prediction. **e**, Summary of the Shapely additive explanations (SHAP)[34] for the Cu(I) sites. **f**, SHAP feature importance for the Cu(II) sites. The colour coding shows the value of the feature (red: high, blue: low). Gray dots show the SHAP values for the other copper site. A negative SHAP value translates into a lower predicted oxidation state, whereas a positive SHAP value corresponds to a higher oxidation state. In all structures, copper is shown in blue, oxygen in red and carbon in brown.

An interesting case of a flexible MOF is MIL-47. For this MOF Barthelet *et al.*[35] reported an oxidation of the V(III) centre upon desorption of a terephtalate guest molecule, which also resulted in a change of flexibility of the framework. In contrast to that, Centrone *et al.*[36] found no evidence for such a change in oxidation state. Our model supports the initial assignment, as also did follow-up studies[37;38]: For the crystal structure with the terephthalate guest molecule (cf. Fig. 6**a**) we find vanadium in the oxidation state +III whereas we find vanadium in the oxidation state +IV for the crystal structures without the guest molecule (cf. Fig. 6**b**). As visible in Fig. 6, the structures show a subtle change in the coordination geometry upon activation, which our model mostly captured in a change of the order parameter for octahedral coordination (coordination number 6 order parameters in Fig. 6**c**–**f**) which is higher for the structure with guest molecule. This reflects the chemical intuition that V(III) is regularly octahedrally coordinated and that the regular octahedron is distorted upon oxidation[39]. It is difficult to capture such subtle effects in deductive approaches like formal electron counting or with the functional form of the bond valence method.

Another peculiar example for the importance of small geometrical details in the assignment of the oxidation states is a redox-active MOF of the MOF-74 type in which the iron centre was shown to be oxidized upon O_2 adsorption at room temperature, which was also reflected in a slight change in the coordination geometry of O_2 from end-on (η^1) to a rather side-on (η^2) coordination[40]. Our model is able to recognise the change in oxidation state based on the slight change in coordination geometry. In a classical bond valence or ligand-counting analysis, the assignment would remain ambiguous due to the dependence on the arbitrary choice of the method to assign bonds between the atoms.

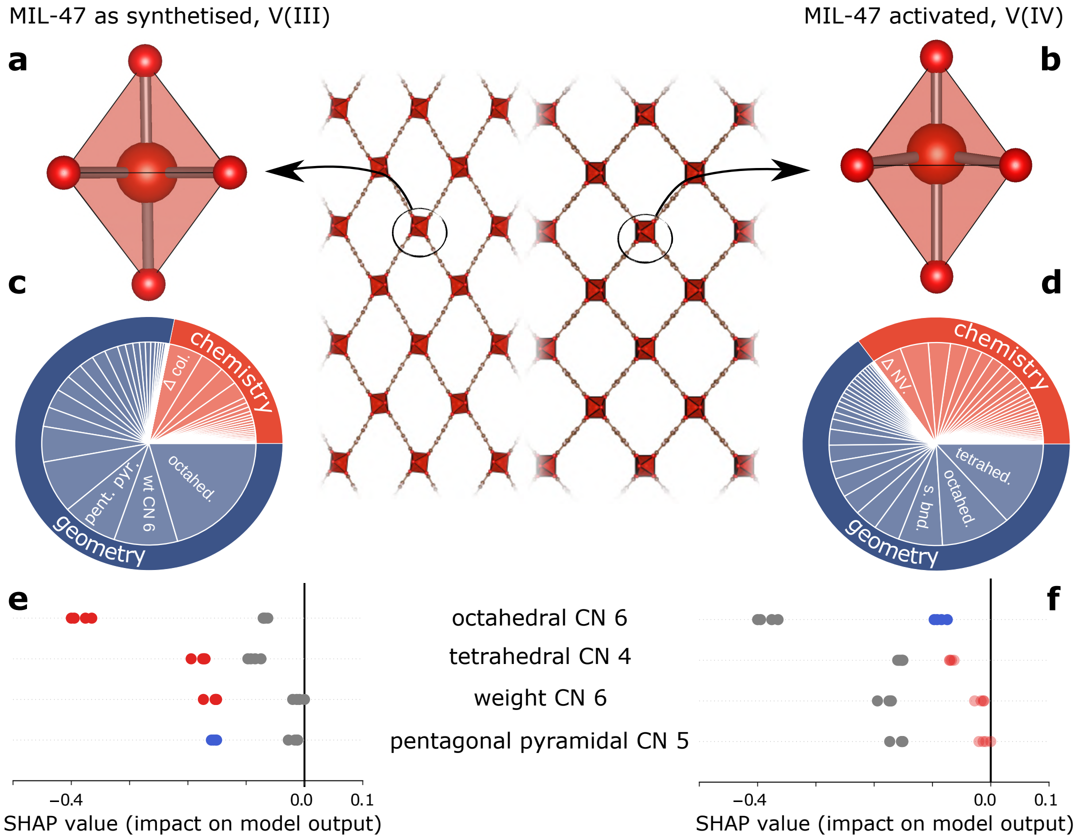

Fig. 6 | Predictions of the oxidation states in MIL-47 before (as synthesised) and after activation. a, Octadedral coordination in the as synthesised structure. **b**, Distortion of the octahedron after activation. **c** and **d**, Global feature importance for the V(II) and V(III) sites, respectively. In both cases, the geometrical features (CN: coordination number) are of the highest importance for the prediction. **e**, Summary of the SHAP feature importance for the V(III) sites. **f**, SHAP feature importance for the V(IV) sites. The colour coding shows the value of the feature (red: high, blue: low). Gray dots show the SHAP values for the other vanadium oxidation state. A negative SHAP value translates into a lower predicted oxidation state, whereas a positive SHAP value corresponds to a higher oxidation state. In all structures, vandadium is shown in orange, oxygen in red.

Novel MOFs

An interesting question is how well we would predict the oxidation state of a novel MOF of which the chemistry is different from structures that are currently in the CSD. One way to estimate the transferability of our model is to test it on databases of other structure classes, including binary ionic solids from the Materials Project[41], transition metal complexes[42] and covalent organic frameworks (COFs)[43], without additional training. For small transition metal complexes, chemists conventionally assign oxidation states by adding up the electron donation of ligands around the metal centre. For ionic crystals on the other hand, chemists will usually base their reasoning directly on the chemical formula. Our model unifies this picture: for the cases in which the model is highly confident in its prediction, we can predict the oxidation state with almost the same accuracy as for MOFs (see Supplementary Information). Moreover, from a more practical point of view, these results give confidence that our approach will predict reasonable oxidation states for novel classes of MOFs that are not yet in the CSD.

We provide an app that uses our pre-trained model to assign oxidation states of metal centres of MOFs on the Materials Cloud. This app requires the crystal structure as input and outputs the oxidation states of the different metal sites together with an estimate of the confidence. In addition, the program can provide details on the feature importance.

Conclusion

Oxidation states are a fundamental concept in chemistry. For many compounds (salts, simple metal complexes) we can write down the oxidation state from empirical knowledge. The bond valence method is successful in assigning oxidation states of more complex structures. For small systems, we can even carry out accurate quantum calculations to determine the oxidation state[44]. However, there are many structures for which these approaches are of limited use. Yet, chemists have provided a large amount of data on the oxidation state of structures for which these conventional approaches cannot be used. In this work, we show that with an appropriate set of descriptors this collective knowledge can be converted into a surprisingly powerful tool. Our work highlights the power data-driven techniques can have in chemistry and materials science; as an example to solve fuzzy problems, where no reliable alternative exists, but relying on the collective knowledge acquired by chemists.

Computational Methods

We used the CSD python application programming interface (API) to retrieve the chemical names for the structures of the MOF subset of May, 2019[45]. Regular expressions were used to parse the oxidation states and the corresponding metals. We excluded 6921 structures from our modeling workflow due to atomic overlaps in the experimental structure.

For featurization, we used the `matminer` python package[46] and standardised (based on standard deviation and mean of each column) all features prior to use in the modelling process.

The ML model adopted in this work is a soft voting classifier using gradient boosting, *k*-nearest neighbours, logistic regression and an extra trees base classifier implemented in the `sklearn` library[47]. For hyperparameter optimisation of each base estimator, we used a mixed strategy of random search, simulated annealing and the tree Parzen estimator (tpe) algorithm for 500 evaluations using the `hyperopt-sklearn` library to avoid biases due to a single search strategy. Classification probabilities were calibrated on a validation set, disjoint from training and test set, using isotonic regression. We use soft voting to be able to provide an uncertainty metric. Further, this approach is appealing as it gives higher weight to more confident models. More details can be found in the Supplementary Information.

To ensure that test errors are not optimistically biased due to multiple similar, but not identical, local environments in one structure we not only constrained the split into training and test set to have the same ratios of oxidation states and elements (iterative stratification[48]) but also to include all chemical environments of one structures in only one set. That is, if one chemical environment of a structure appears in the training set all other chemical environments of the same structure will not appear in the test set. Identical fingerprints are automatically discarded from our training set. We perform this split based on "base identifiers" of the CSD database identifiers, which we create by stripping all trailing integers. This accounts for the fact that some entries in the CSD are updated entries (e.g., with refined lattice constants) for the same structure for which a trailing number has been added to the original identifier. By restricting all structures with the same base identifier to be in the same set, we avoid data leakage.

Further, we use a submodular selection approach[49] to select a smaller, diverse set of training points to make our training more efficient (and again recognise the parsimony principle of Pauling by minimizing redundancy in our training set). To address the fact that some metals (like copper) are more than an order of magnitude more frequent than other metals (like ruthenium) we adjusted our sampling procedure to randomly subsample the structures with the most common metals (Cu, Zn, Cd).

Crystal structures were drawn using VESTA[50].

Acknowledgments

This work is supported by the European Research Council (ERC) Advanced Grant (Grant Agreement No.666983, MaGic) and the computing facilities of SCITAS, EPFL. The authors want to thank Leopold Talirz and the Materials Cloud team for feedback on the code for the web app.

Author contributions

K.M.J developed the machine learning workflows, D.O. carried out the bond valence sum analysis. B.S., S.M.M, and K.M.J developed the featurization. All authors contributed to the analysis of the data and the writing of the article.

Competing interest

The authors declare no competing interests.

Data availability

The feature matrices, labels and a pre-trained model are deposited on the Materials Cloud archive (DOI: 10.24435/materialscloud:2019.0085/v1).

Code availability

The code for parsing, featurization as well for the ML models is available on Github (https://github.com/kjappelbaum/learn_mof_ox_state/tree/master, https://github.com/kjappelbaum/mof_oxidation_states) and deposited on Zenodo (DOIs: 10.5281/zenodo.3567011, 10.5281/zenodo.3567274). The web app is hosted on the work section of Materials Cloud (https://dev-tools.materialscloud.org/oximachine). The code for this app is also available on Github (https://github.com/kjappelbaum/oximachinetool) and deposited on Zenodo (DOI: 10.5281/zenodo.3603606. All codes are available under GNU General Public License v3.

References

1. Jensen, W. B. The Origin of the Oxidation-State Concept. *J. Chem. Educ.* **84**, 1418 (2007).

2. Wöhler, F. Oxydation, Oxyde. In *Grundriss Der Chemie: Unorganische Chemie*, 3 (Duncker und Humblot, Berlin, 1835).

3. Latimer, W. M. *The Oxidation States of the Elements and Their Potentials in Aqueous Solutions*. Prentice-Hall Chemistry Series (Prentice-Hall, Englewood Cliffs, 1952), second edition edn.

4. Connelly, N. G., Royal Society of Chemistry (Great Britain) & International Union of Pure and Applied Chemistry (eds.) *Nomenclature of Inorganic Chemistry. IUPAC Recommendations 2005* (Royal Society of Chemistry Publishing/IUPAC, Cambridge, UK, 2005). OCLC: ocm60838140.

5. Kroll, J. H. *et al.* Carbon oxidation state as a metric for describing the chemistry of atmospheric organic aerosol. *Nature Chem* **3**, 133–139 (2011).

6. Terrett, J. A., Cuthbertson, J. D., Shurtleff, V. W. & MacMillan, D. W. C. Switching on elusive organometallic mechanisms with photoredox catalysis. *Nature* **524**, 330–334 (2015).

7. Jørgensen, C. K. *Oxidation Numbers and Oxidation States* (Springer Berlin Heidelberg, Berlin, Heidelberg, 1969).

8. Bendix, J., Brorson, M. & Schäffer, C. E. Oxidation States and d ^{q} Configurations in Inorganic Chemistry: A Historical and Up-to-Date Account. In Kauffman, G. B. (ed.) *Coordination Chemistry*, vol. 565, 213–225 (American Chemical Society, Washington, DC, 1994).

9. Nič, M., Jirát, J., Košata, B., Jenkins, A. & McNaught, A. (eds.) *IUPAC Compendium of Chemical Terminology: Gold Book* (IUPAC, Research Triagle Park, NC, 2009), 2.1.0 edn.

10. Karen, P., McArdle, P. & Takats, J. Comprehensive definition of oxidation state (IUPAC Recommendations 2016). *Pure and Applied Chemistry* **88**, 831–839 (2016).

11. Walsh, A., Sokol, A. A., Buckeridge, J., Scanlon, D. O. & Catlow, C. R. A. Electron Counting in Solids: Oxidation States, Partial Charges, and Ionicity. *J. Phys. Chem. Lett.* **8**, 2074–2075 (2017). ZSCC: 0000024.

12. Brown, I. D. Recent Developments in the Methods and Applications of the Bond Valence Model. *Chem. Rev.* **109**, 6858–6919 (2009).

13. Pauling, L. Atomic Radii and Interatomic Distances in Metals. *J. Am. Chem. Soc.* **69**, 542–553 (1947).

14. Shields, G. P., Raithby, P. R., Allen, F. H. & Motherwell, W. D. S. The assignment and validation of metal oxidation states in the Cambridge Structural Database. *Acta Crystallogr B Struct Sci* **56**, 455–465 (2000).

15. Taylor, R. & Wood, P. A. A Million Crystal Structures: The Whole Is Greater than the Sum of Its Parts. *Chem. Rev.* **119**, 9427–9477 (2019).

16. O'Keeffe, M. A proposed rigorous definition of coordination number. *Acta Crystallogr A Cryst Phys Diffr Theor Gen Crystallogr* **35**, 772–775 (1979).

17. Brown, I. A determination of the oxidation states and internal stresses in Ba2YCu3Ox, x = 6 – 7 using bond valences. *J. Solid State Chem.* **82**, 122–131 (1989).

18. Müller, P., Köpke, S. & Sheldrick, G. M. Is the bond-valence method able to identify metal atoms in protein structures? *Acta Crystallogr D Biol Crystallogr* **59**, 32–37 (2003).

19. Gagné, O. C. Bond-length distributions for ions bonded to oxygen: Results for the lanthanides and actinides and discussion of the *f* -block contraction. *Acta Crystallogr B Struct Sci Cryst Eng Mater* **74**, 49–62 (2018).

20. Wang, L., Maxisch, T. & Ceder, G. Oxidation energies of transition metal oxides within the GGA + U framework. *Phys. Rev. B* **73**, 195107 (2006).

21. Stevanović, V., Lany, S., Zhang, X. & Zunger, A. Correcting density functional theory for accurate predictions of compound enthalpies of formation: Fitted elemental-phase reference energies. *Phys. Rev. B* **85**, 115104 (2012).

22. Raebiger, H., Lany, S. & Zunger, A. Charge self-regulation upon changing the oxidation state of transition metals in insulators. *Nature* **453**, 763–766 (2008).

23. Groom, C. R., Bruno, I. J., Lightfoot, M. P. & Ward, S. C. The Cambridge Structural Database. *Acta Crystallogr B Struct Sci Cryst Eng Mater* **72**, 171–179 (2016).

24. Holgate, S. CSD Data Curation – The Human Touch - The Cambridge Crystallographic Data Centre (CCDC). https://www.ccdc.cam.ac.uk/Community/blog/CSD-data-curation-the-human-touch/ (2019).

25. Huang, B. & von Lilienfeld, O. A. Communication: Understanding molecular representations in machine learning: The role of uniqueness and target similarity. *J. Chem. Phys.* **145**, 161102 (2016).

26. Ghiringhelli, L. M., Vybiral, J., Levchenko, S. V., Draxl, C. & Scheffler, M. Big Data of Materials Science: Critical Role of the Descriptor. *Phys. Rev. Lett.* **114**, 105503 (2015).

27. Pauling, L. THE PRINCIPLES DETERMINING THE STRUCTURE OF COMPLEX IONIC CRYSTALS. *J. Am. Chem. Soc.* **51**, 1010–1026 (1929).

28. Prodan, E. & Kohn, W. Nearsightedness of electronic matter. *Proc Natl Acad Sci U S A* **102**, 11635–11638 (2005).

29. Ward, L. *et al.* Including crystal structure attributes in machine learning models of formation energies via Voronoi tessellations. *Physical Review B* **96** (2017).

30. Rokach, L. Ensemble-based classifiers. *Artif Intell Rev* **33**, 1–39 (2010).

31. Nouar, F. *et al.* Tuning the properties of the UiO-66 metal organic framework by Ce substitution. *Chem. Commun.* **51**, 14458–14461 (2015).

32. Stawowy, M. *et al.* The Impact of Synthesis Method on the Properties and CO2 Sorption Capacity of UiO-66(Ce). *Catalysts* **9**, 309 (2019).

33. Ahmed, A. *et al.* Cu(I)Cu(II)BTC, a microporous mixed-valence MOF via reduction of HKUST-1. *RSC Adv.* **6**, 8902–8905 (2016).

34. Lundberg, S. M. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. In Guyon, I. *et al.* (eds.) *Advances in Neural Information Processing Systems 30*, 4765–4774 (Curran Associates, Inc., 2017).

35. Barthelet, K., Marrot, J., Riou, D. & Férey, G. A Breathing Hybrid Organic–Inorganic Solid with Very Large Pores and High Magnetic Characteristics. *Angew. Chem. Int. Ed.* **41**, 281–284 (2002).

36. Centrone, A., Harada, T., Speakman, S. & Hatton, T. A. Facile Synthesis of Vanadium Metal-Organic Frameworks and their Magnetic Properties. *Small* **6**, 1598–1602 (2010).

37. Leclerc, H. *et al.* Influence of the Oxidation State of the Metal Center on the Flexibility and Adsorption Properties of a Porous Metal Organic Framework: MIL-47(V). *J. Phys. Chem. C* **115**, 19828–19840 (2011).

38. Kozachuk, O. *et al.* A Solid-Solution Approach to Mixed-Metal Metal-Organic Frameworks - Detailed Characterization of Local Structures, Defects and Breathing Behaviour of Al/V Frameworks. *Eur. J. Inorg. Chem.* **2013**, 4546–4557 (2013).

39. Krakowiak, J., Lundberg, D. & Persson, I. A Coordination Chemistry Study of Hydrated and Solvated Cationic Vanadium Ions in Oxidation States +III, +IV, and +V in Solution and Solid State. *Inorg. Chem.* **51**, 9598–9609 (2012).

40. Bloch, E. D. *et al.* Selective Binding of O_2 over N_2 in a Redox–Active Metal–Organic Framework with Open Iron(II) Coordination Sites. *J. Am. Chem. Soc.* **133**, 14814–14822 (2011).

41. Jain, A. *et al.* Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials* **1**, 011002 (2013).

42. Janet, J. P. & Kulik, H. J. Predicting electronic structure properties of transition metal complexes with neural networks. *Chem. Sci.* **8**, 5137–5152 (2017).

43. Ongari, D., Yakutovich, A. V., Talirz, L. & Smit, B. Building a Consistent and Reproducible Database for Adsorption Evaluation in Covalent–Organic Frameworks. *ACS Cent. Sci.* **5**, 1663–1675 (2019).

44. Jiang, L., Levchenko, S. V. & Rappe, A. M. Rigorous Definition of Oxidation States of Ions in Solids. *Phys. Rev. Lett.* **108**, 166403 (2012).

45. Moghadam, P. Z. *et al.* Development of a Cambridge Structural Database Subset: A Collection of Metal–Organic Frameworks for Past, Present, and Future. *Chem. Mater.* **29**, 2618–2625 (2017).

46. Ward, L. *et al.* Matminer: An open source toolkit for materials data mining. *Computational Materials Science* **152**, 60–69 (2018).

47. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).

48. Sechidis, K., Tsoumakas, G. & Vlahavas, I. On the Stratification of Multi-label Data. In Gunopulos, D., Hofmann, T., Malerba, D. & Vazirgiannis, M. (eds.) *Machine Learning and Knowledge Discovery in Databases*, vol. 6913, 145–158 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2011).

49. Schreiber, J., Bilmes, J. & Noble, W. S. Apricot: Submodular selection for data summarization in Python. *arXiv:1906.03543* (2019). `1906.03543`.

50. Momma, K. & Izumi, F. VESTA 3 for three-dimensional visualization of crystal, volumetric and morphology data. *J Appl Cryst* **44**, 1272–1276 (2011).