Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *Digital Discovery*, 2024, **3**, 1638

Automated prediction of ground state spin for transition metal complexes†

Yuri Cho, [ab] Ruben Laplaza, [ac] Sergi Vela [de] and Clémence Corminboeuf *[abc]

Exploiting crystallographic data repositories for large-scale quantum chemical computations requires the rapid and accurate extraction of the molecular structure, charge and spin from the crystallographic information file. Here, we develop a general approach to assign the ground state spin of transition metal complexes, in complement to our previous efforts on determining metal oxidation states and bond order within the *cell2mol* software. Starting from a database of 31k transition metal complexes extracted from the Cambridge Structural Database with *cell2mol*, we construct the TM-GSspin dataset, which contains 2063 mononuclear first row transition metal complexes and their computed ground state spins. TM-GSspin is highly diverse in terms of metals, metal oxidation states, coordination geometries, and coordination sphere compositions. Based on TM-GSspin, we identify correlations between structural and electronic features of the complexes and their ground state spins to develop a rule-based spin state assignment model. Leveraging this knowledge, we construct interpretable descriptors and build a statistical model achieving 98% cross-validated accuracy in predicting the ground state spin across the board. Our approach provides a practical way to determine the ground state spin of transition metal complexes directly from crystal structures without additional computations, thus enabling the automated use of crystallographic data for large-scale computations involving transition metal complexes.

Received 8th April 2024
Accepted 10th July 2024

DOI: 10.1039/d4dd00093e

rsc.li/digitaldiscovery

1 Introduction

The automated construction of datasets has become increasingly relevant for data-driven computational chemistry.[1–3] Data-driven approaches to computational chemistry essentially include high-throughput screening of molecules and materials by quantum chemical (QC) computations[4–9] as well as the use of large-scale computed data to train machine learning (ML) models for property prediction.[10–19] Both tasks rely on extensive datasets curated to cover vast and diverse regions of chemical space.[20–22] Within this context, crystallographic data repositories constitute a valuable pool of synthesized structures available in large size and chemical diversity.[23–26] The Cambridge Structural Database (CSD)[27,28] contains, for instance, over a million of experimental crystal structures collected over several decades.

Yet, the proper exploitation of crystallographic information for the computational chemistry is not straightforward. For datasets to be adapted for both the training of ML models and high-throughput QC searches, they must include the essential information needed to run an electronic structure computation, such as the structure (R), the molecular charge (Q) as well as the spin multiplicity. Owing to the lack of information about metal oxidation (OS)[29] and spin states, reliably retrieving the molecular charge and spin multiplicity is especially difficult when transition metals (TM) are involved. To overcome this limitation, we recently developed *cell2mol*,[22] a software that specifically interprets crystallographic data to retrieve the Cartesian coordinates, total charges, and connectivity of all individual molecules in the unit cell, including the OS of metal ions. While *cell2mol* provides a thorough unit cell interpretation, the original version was not coded to characterize ground state spins.

Given that the ground state spin of TM complexes depends on multiple factors, such as metal identity, OS, coordination geometry and ligand field strength, deducing this information only from their structure is challenging.[30–32] Significant efforts have been made to train ML models that predict spin-state-dependent properties such as spin-splitting energies, spin-state orderings, sensitivity to Hartree–Fock exchange, and metal–ligand bond lengths[25,33–35] but these efforts have been essentially placed on mononuclear octahedral complexes and

[a]*Laboratory for Computational Molecular Design, Institute of Chemical Sciences and Engineering, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland. E-mail: clemence.corminboeuf@epfl.ch*

[b]*National Centre for Computational Design and Discovery of Novel Materials (MARVEL), École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland*

[c]*National Centre for Competence in Research-Catalysis (NCCR-Catalysis), École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland*

[d]*Departament de Ciència de Materials i Química Física and IQTCUB, Universitat de Barcelona, Barcelona, Spain*

[e]*Institut de Química Avançada de Catalunya (IQAC-CSIC), Barcelona, Spain*

† Electronic supplementary information (ESI) available. See DOI: https://doi.org/10.1039/d4dd00093e

on a restricted range of exemplary ligands with varying field strengths along the spectrochemical series. So far, the prediction of ground state spin of TM complexes has not been investigated across diverse chemical spaces.

Herein, we develop a pragmatic and general workflow (Fig. 1) to predict the ground state spin of TM complexes by leveraging previously curated data obtained with *cell2mol*. Starting from the original database that was extracted from the CSD, we construct a smaller, representative, albeit diverse set through stratified sampling. We then determine the ground state spin of each individual complex through density functional theory (DFT) computation using the B3LYP* functional[36,37] and filter out ambiguous cases. The resulting TM-GSspin dataset is systematically analyzed to identify correlations between the structural and electronic features of the complexes and their ground state spins. Based on the extracted patterns and relationships, we construct rule-based empirical and interpretative random forest models for ground state spin assignment in first row TM complexes. These models are integrated into *cell2mol*, enabling the assignment of total charge, OS, and ground state spin of TM complexes directly from crystallographic information files.

2 Methods

2.1 Dataset generation

Emphasis is placed on first row TM complexes with d electron configurations ranging from d^4 to d^8 that adopt different spin states depending on the nature of the metal and its coordination environment. Note that second and third row TM complexes are less cumbersome, as they typically exhibit low spin configurations due to larger crystal field splitting.[38,39] As a starting point, we took the database containing 31k transition metal complexes we extracted from the CSD (updated May 2021) using *cell2mol* version 1.1.0.[22] This database[22] excludes polynuclear complexes, for which the total spin assignment would depend upon the coupling between spin-bearing metal centers, and complexes with formally radical ligands.

This results in 17 214 mononuclear first row complexes with five metal centers: Cr, Mn, Fe, Co, and Ni. Among these complexes, we excluded those with haptic ligands (*e.g.*, cyclopentadienyl) because their coordination numbers and geometries are ambiguous, presenting subtly different η coordination modes.[40] We also eliminated complexes with nitrosyl ligands to avoid potential spin from typical non-innocent ligands.[41] The coordination geometry of the TM complex was then determined using the CoSymLib python library[42] and complexes exhibiting significant deviation from the ideal shape of a reference polyhedron were removed to unequivocally identify the correlation between the ground state spin and the coordination geometry (see Section S1 and Fig. S1 in the ESI†).

The remaining 15 837 complexes were classified based on metal identity, OS, coordination number (the number of atoms bound to the metal center), coordination geometry, and composition of the metal-coordinating atoms, resulting in 1633 distinct groups of complexes that share the aforementioned characteristics. We then performed stratified sampling among those groups to construct a dataset of 2261 complexes where each group is represented (see Section S1 in the ESI† for further details regarding dataset construction and curation, and Section S2† for a discussion on the excluded complexes).

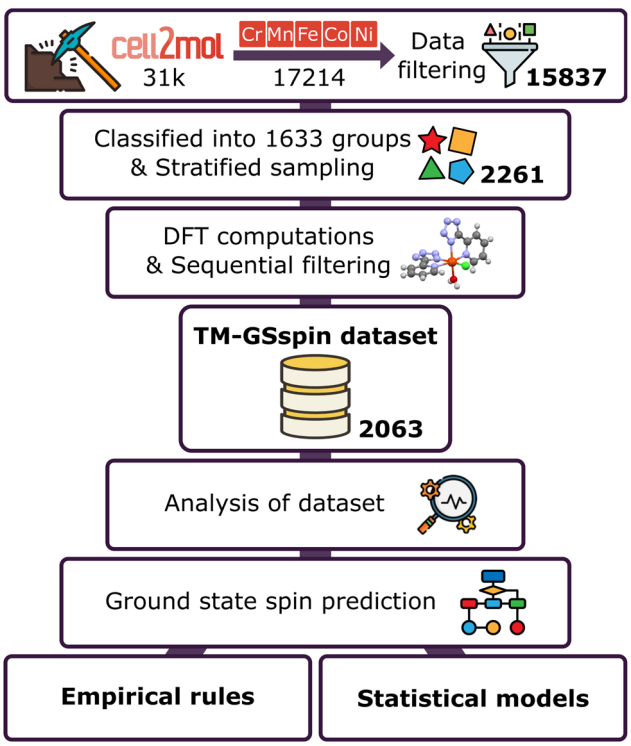

Fig. 1 Proposed general workflow. The numbers in bold indicate the number of complexes curated at each step.

2.2 Ground state spin computations

Crystal structure geometries were refined by optimizing the atomic positions of hydrogen atoms in either the singlet or doublet state, as hydrogen atoms generally exhibit the greatest uncertainty in refinement due to their small electron density. Optimizations were carried out using Gaussian09 (revision D.01)[43] at the B3LYP*-D3(BJ)/def2-SVP level.[44,45] B3LYP* is a reparametrized version of B3LYP that reduces Hartree–Fock exchange from 20% to 15%, and was shown to improve the description of spin-splitting energetics in TM complexes.[36,37] Single point computations were then performed at the B3LYP*-D3(BJ)/def2-TZVP level for three different spin states: singlet, triplet, and quintet for systems with an even number of electrons, and doublet, quartet, and sextet otherwise. The spin state with the lowest energy was assigned as the ground state spin. B3LYP* results were compared to two other functionals, TPSSh and M06L, which have also demonstrated good performance in describing splitting energies in spin-crossover Fe complexes.[46,47] Except for sensitive cases (*vide infra*), consistent ground state spins were observed with all three levels (see Fig. S3 in the ESI†).

An iterative and automated correction was applied in case of convergence failures. Complexes for which computations failed to converge in all accessible spin states were removed. Those

with energy gaps between possible spin states smaller than 5 kcal mol^{-1} were excluded, as they fall below the chemical accuracy of DFT for spin-state energetics of TM complexes.[48–50] This filtering step also eliminates spin-crossover complexes, for which ground spin state changes with an external stimulus like temperature or pressure as a result of vibrational and electronic entropy contributions.[33,46,51,52] Finally, complexes exhibiting an expectation value of $\langle \hat{S}^2 \rangle$ that deviates from the exact value of $S(S+1)$ by more than 0.1 for the singlet and doublet ground states, and more than 0.2 for the other ground state spins[49] were also excluded. Overall, a total of 198 complexes were removed by the three consecutive filters. For further details on the in-depth analysis of the DFT results, excluded complexes, impact of geometry optimization on spin-splitting energies, and complexes with hydride ligands, see Section S2 in the ESI.†

3 Results and discussion

3.1 TM-GSspin dataset

The final curated TM-GSspin dataset consists of 2063 mononuclear complexes and their corresponding ground state spins. The following subsections analyze the dataset and more specifically the relationship between structural and electronic features of the complexes and their ground state spins.

3.1.1 Chemical diversity of the dataset. Fig. 2 illustrates the large chemical diversity of the TM-GSspin dataset, encompassing various metal identities, OS, coordination geometries, as well as compositions of the first coordination sphere. Each metal exhibits three OSs, with a number of d electrons ranging from 3 to 8 (Fig. 2a). Eighteen types of coordination geometries are obtained, with coordination numbers ranging from 2 to 8 (Fig. 2b). Those include common geometries like octahedral, tetrahedral, or square planar, as well as less common ones such as linear, trigonal planar, pentagonal bipyramidal, or capped trigonal prismatic. Nitrogen is found to be the most recurrent metal-coordinating atoms followed by oxygen, carbon, sulfur, phosphorus, chlorine, and bromine (Fig. 2c). Combinations of different metal-coordinating elements across various coordination numbers and geometries lead to over 600 different first coordination spheres, including complexes with up to 5 different coordinating elements (Fig. 2d). In comparison with the original database[22] of 15 837 complexes mined from the CSD, the TM-GSspin dataset is characterized by an increased proportion of complexes with less common coordination geometries (*e.g.*, the proportion of trigonal planar complexes increases from 0.7% in the original database to 2.7% in the TM-GSspin dataset) and those with zero or low-valent OSs (see comparison in Fig. S2 in the ESI†). Additionally, Fig. S6 and S7 in the ESI† provide an overview of the TM-GSspin dataset, including the number of atoms, number of electrons, total molecular charges, and spin multiplicity, as well as the sizes of the ligands.

3.1.2 Analysis of the transition metal complexes ground state spins. Fig. 3a shows the proportion of ground state spin of TM complexes for each metal and OS. Some metals in a given OS consistently exhibit the same ground state spin. As expected, if the number of d electrons is less than 4 or more than 8, the

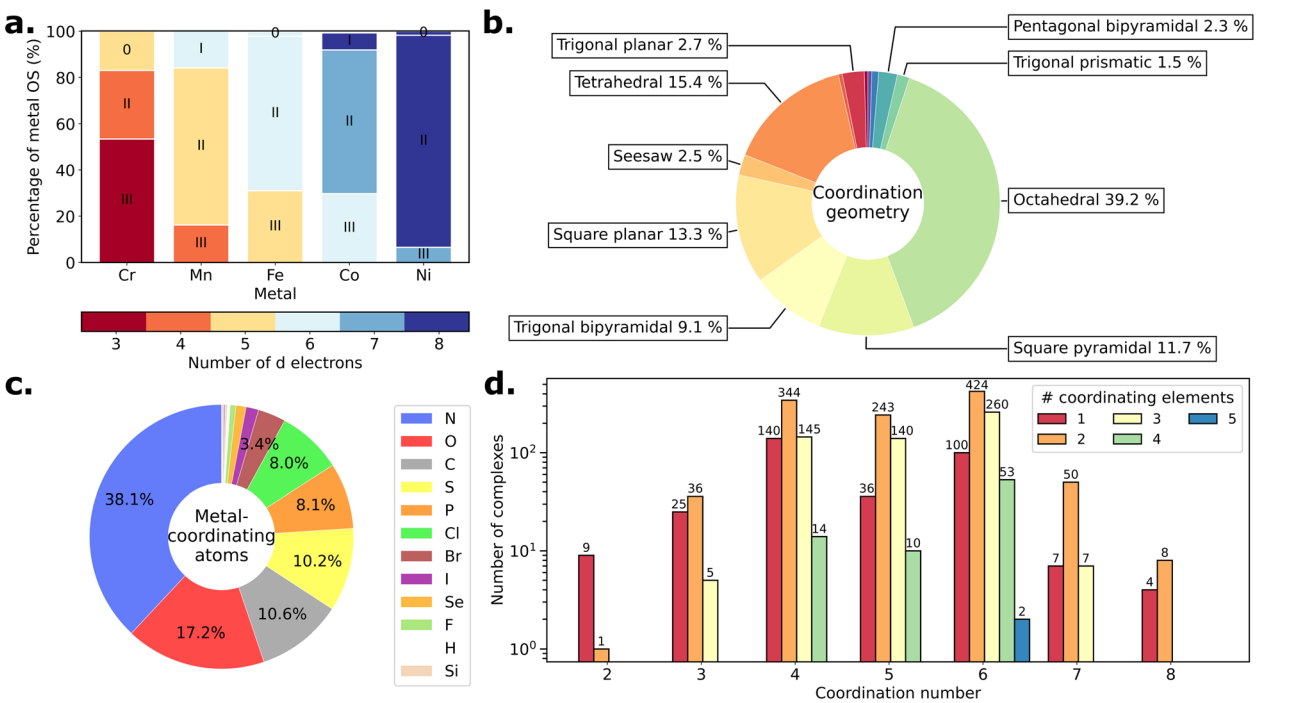

Fig. 2 Chemical diversity of the TM-GSspin dataset. (a) Percentage of oxidation state (OS) for each metal. The OS is denoted 0, I, II, or III. The color code indicates the number of d electrons in the metal ion. (b) Frequency distribution of coordination geometries. (c) Frequency distribution of metal-coordinating atom elemental identities. Elements with a frequency < 0.2% are omitted in the legend box. (d) Number of different elements in the first coordination sphere for each coordination number.

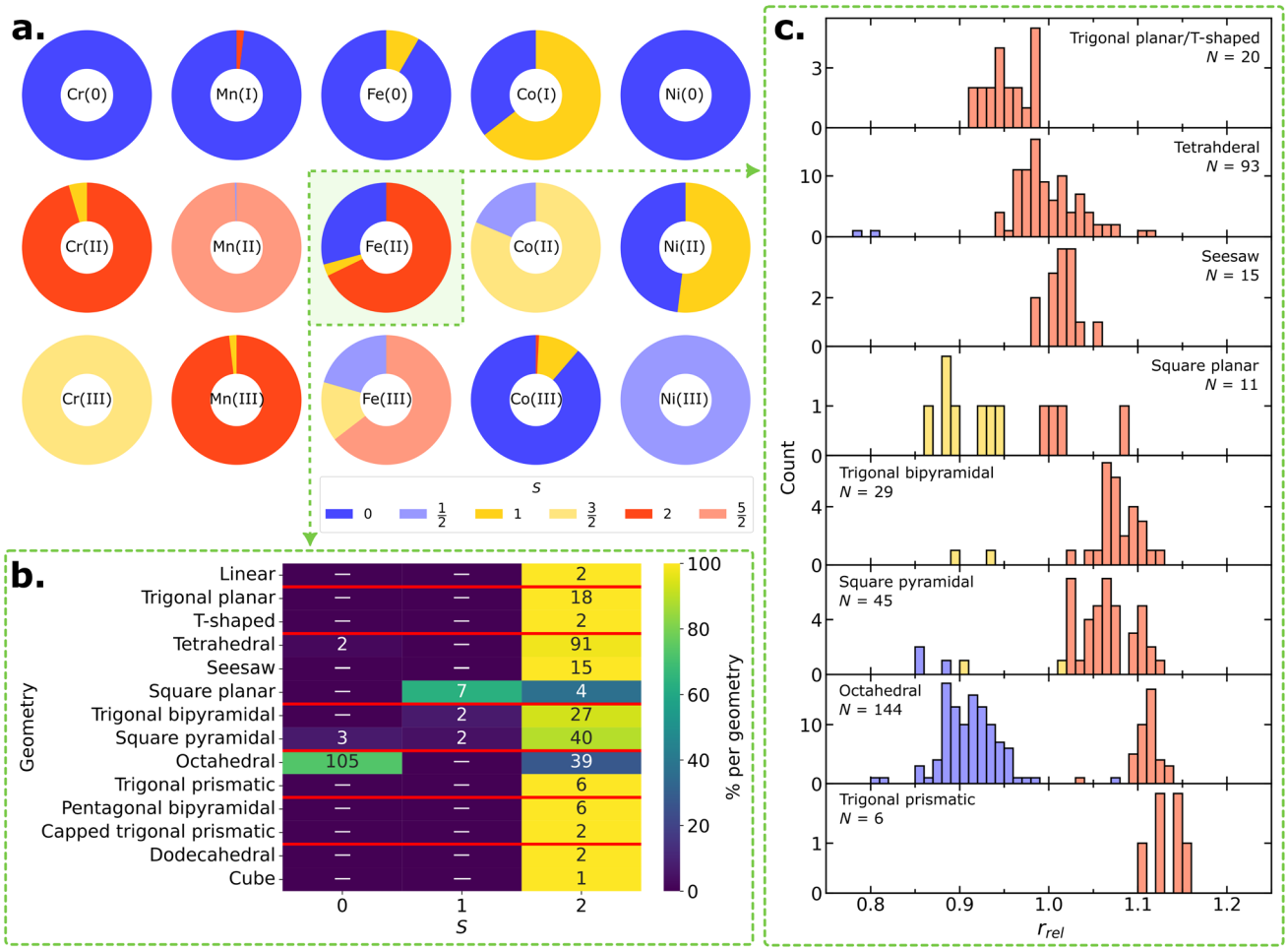

Fig. 3 Relationship between ground state spins and various features of TM complexes. (a) Proportion of ground state spins for different metal centers and oxidation states. (b) Ground state spins of d^6 Fe(II) complexes based on their coordination geometries. Each column represents a total spin quantum number *S*, and each row represents a coordination geometry. Red horizontal lines classify coordination geometries by their coordination numbers. The number within each grid cell indicates the number of corresponding complexes in that cell (0 is shown as a hyphen). The color code represents the proportion of spin states within a given coordination geometry, where navy blue corresponds to 0% and yellow to 100%. (c) Histograms of relative metal radii (r_{rel}) and corresponding ground state spins for Fe(II) complexes with coordination numbers ranging from 3 to 6. *N* indicates the number of complexes used to plot each histogram. The color code represents the ground state spin: blue (singlet), yellow (triplet), and red (quintet).

number of unpaired d electrons is the determining factor. For instance, all Cr(III) complexes (d^3 configuration) have a quartet ground state. Alternatively, zero–valent complexes such as Cr(0), Fe(0), or Ni(0) centers as well as and Mn(I) complexes possess a singlet ground state. These complexes typically bind to strong-field ligands such as carbonyls or substituted phosphines, which cause a substantial energy separation between d orbitals and favor lower spin states. Intriguingly, one Fe(0) tetrahedral complex (CSD refcode: NUNWUP[53]) and one Mn(I) linear complex (CSD refcode: CUJSAD[54]) deviate from this trend (Fig. S8 in the ESI†). The behavior of the former can be attributed to eight valence electrons in the tetrahedral crystal field, while the linearity in the latter, reported as $[K(15\text{-crown-}5)_2][Mn\{C(SiMe_3)_2\}]$, is affected by crystal packing effects.[54] The d^7 Ni(III) complexes, in the dataset, constitute another constant example that exclusively adopt the doublet ground state.

TM centers exhibiting various ground state spins can be understood based on their coordination environments. As an illustrative example, the analysis of the ground state spins of Fe(II) complexes across fourteen different coordination geometries is provided in Fig. 3b d^6 Fe(II) complexes adopt singlet, triplet, or quintet ground state, which corresponds to low-spin (LS), intermediate-spin (IS), or high-spin (HS) state, respectively. Most coordination geometries of these complexes exhibit the HS ground state. Specifically, all Fe(II) complexes with coordination numbers smaller than 4 or greater than 6 (*i.e.*, linear, trigonal planar, T-shaped, pentagonal bipyramidal, capped trigonal prismatic, dodecahedral and cube geometries) are consistently HS. Tetrahedral, seesaw, and trigonal prismatic Fe(II) complexes are all HS except for two tetrahedral imido complexes containing tertiary phosphine ligands, which adopt LS states. Fe(II) complexes with other coordination geometries exhibit a greater ground state spin variability. Square planar

and trigonal bipyramidal Fe(II) complexes exhibit IS or HS, depending on whether the highest d orbital is empty or half-filled. Square pyramidal Fe(II) complexes cover three different ground state spins, while most common octahedral Fe(II) complexes adopt either the LS or HS ground states.

Similar trends are observed for other TM complexes with d^4 to d^8 electron configurations (see in Fig. S9–S13 of the ESI†). Complexes with low- or high-coordination geometries tend to favor the HS state within a given d electron configuration. Certain coordination geometries, such as tetrahedral and trigonal prismatic, exclusively adopt the HS states due to the small d-orbital splitting in these crystal fields. However, a few tetrahedral complexes favor the LS ground state owing to the presence of strong-field ligands such as substituted phosphines. Complexes with other coordination geometries exhibit different ground state spins depending on the arrangement of d electrons within a given crystal field, which indicates that considering additional factors is crucial for determining the ground state spin. For further analysis, the relationship between ground state spins and coordination sphere compositions of 144 Fe(II) octahedral complexes are shown in Fig. S14 in the ESI,† with a brief discussion.

We further examine the distribution of distances between the metal center and its coordinating atoms as the metal–ligand bond lengths in HS states are generally longer compared to those in the LS states due to the population of anti-bonding orbitals. Within this context, Taylor *et al.*,[25] assigned ground state spins of mononuclear octahedral Fe(II)/Fe(III) complexes based on heuristic cut-off values for metal–ligand bond lengths. We here introduce a more general indicator applicable to various metal centers and coordination geometries. We define the relative metal radius, denoted as r_{rel}, as

$$r_{\mathrm{rel}} = \frac{1}{r_{\mathrm{M}}} \times \frac{\sum_{i=1}^{\mathrm{CN}} \mathrm{d}(\mathrm{M}-\mathrm{A}_i) - r_{\mathrm{A}_i}}{\mathrm{CN}} \quad (1)$$

where CN is the coordination number of a given complex, $\mathrm{d}(\mathrm{M}-\mathrm{A}_i)$ represents the distance between a metal center (M) and a coordinating atom (A_i), and r_{M} and r_{A_i} are covalent radii of M and A_i, respectively. Covalent radii values were taken from a previous analysis of experimental crystal structures (Table S8 in the ESI†).[55]

Fig. 3c shows the distribution of relative metal radii and their corresponding ground state spins for Fe(II) complexes depending on the coordination geometry (see Fig. S15 in the ESI† for Fe(III) complexes). Across different coordination geometries, Fe(II) complexes in the HS state typically possess longer relative metal radii compared to those in the LS or IS states. In general, we observe a biased distribution toward HS states with longer relative metal radii. In particular, the relative metal radii of octahedral Fe(II) complexes show a binomial distribution, separating LS and HS. Interestingly, two tetrahedral Fe(II) complexes with LS state exhibit very short relative metal radii, which deviate from the overall distribution of relative metal radii in tetrahedral complexes. This suggests that relative metal radius serves to identify outliers exhibiting uncommon ground state spins. For further analysis, we investigate one octahedral singlet complex (CSD refcode: DOQRAC) with longer relative metal radii in Fig. 3c, which was identified as a spin-crossover complex in the literature.[56] Moreover, there is a systematic increase in relative metal radius as the coordination number increases. This pattern is especially evident in HS complexes, which display a linear increase, as shown in Fig. S16 in the ESI.†

3.2 Ground state spin assignment based on empirical rules

Based on the trends and relationships observed in the TM-GSspin dataset, we develop an empirical model to assign the most probable ground state spin for first row TM complexes. Fig. 4 shows a rule-based decision tree used in our assignment model, which systematically considers key structural and electronic features of the complexes. The decision tree begins by considering two key factors: the number of d electrons and the OS of the metal center. For coordination complexes with 0, 1, 2, 3, 9, or 10 d electrons, the ground state spin is determined based on the number of unpaired electrons. For complexes with d electrons between 4 and 8 (excluding zero-valent complexes), the decision tree takes into account the coordination environments. Note that d^6 Co(III) octahedral (Fig. S11†) and d^7 Ni(III) complexes (Fig. S12†) exclusively exhibit the LS state, regardless of their coordination environments. Therefore, their ground state spins are assigned as singlet and doublet, respectively.

For complexes with a coordination number of 2 or greater than 6, the ground state spin is assigned as HS based on observed trends. We hypothesize that in such low- or high-coordination cases, the ligand field is weak due to the limited interaction between the metal and ligands. This stems from either sterically hindered bulky ligands in the former case (Table S9 in the ESI†) or overly crowded coordination spheres in the latter. Accordingly, both extremes favor the HS ground state.

For the remaining complexes, the assignment depends on their coordination geometry and relative metal radius. For cases with multiple ground state spins within a given coordination geometry, the model uses the relative metal radius as a distinguishing criterion (indicated by the orange rhombus in Fig. 4). For this step, we define a specific cut-off for each combination of metal and coordination geometry (Table S10 in the ESI†). If the relative metal radius falls below the designated cut-off value, the ground state spin is assigned as the lowest possible spin state. Note that for d^4 octahedral as well as d^6 square planar or trigonal bipyramidal complexes, the lowest spin state is assigned as triplet based on the pattern discerned in the dataset.

Despite its relative simplicity, the empirical model achieves a high 97% accuracy within the dataset. Out of the 2063 complexes considered, only 55 exhibit discrepancies in their ground state spin assignments with respect to the computations. Most of the disagreements occur for square pyramidal complexes featuring Fe(III), Co(II), and Ni(II) with relative metal radii close to the cut-off values. Furthermore, our empirical model is indeed unable to assign the IS state to square pyramidal Fe complexes, which would require an additional cut-off value.

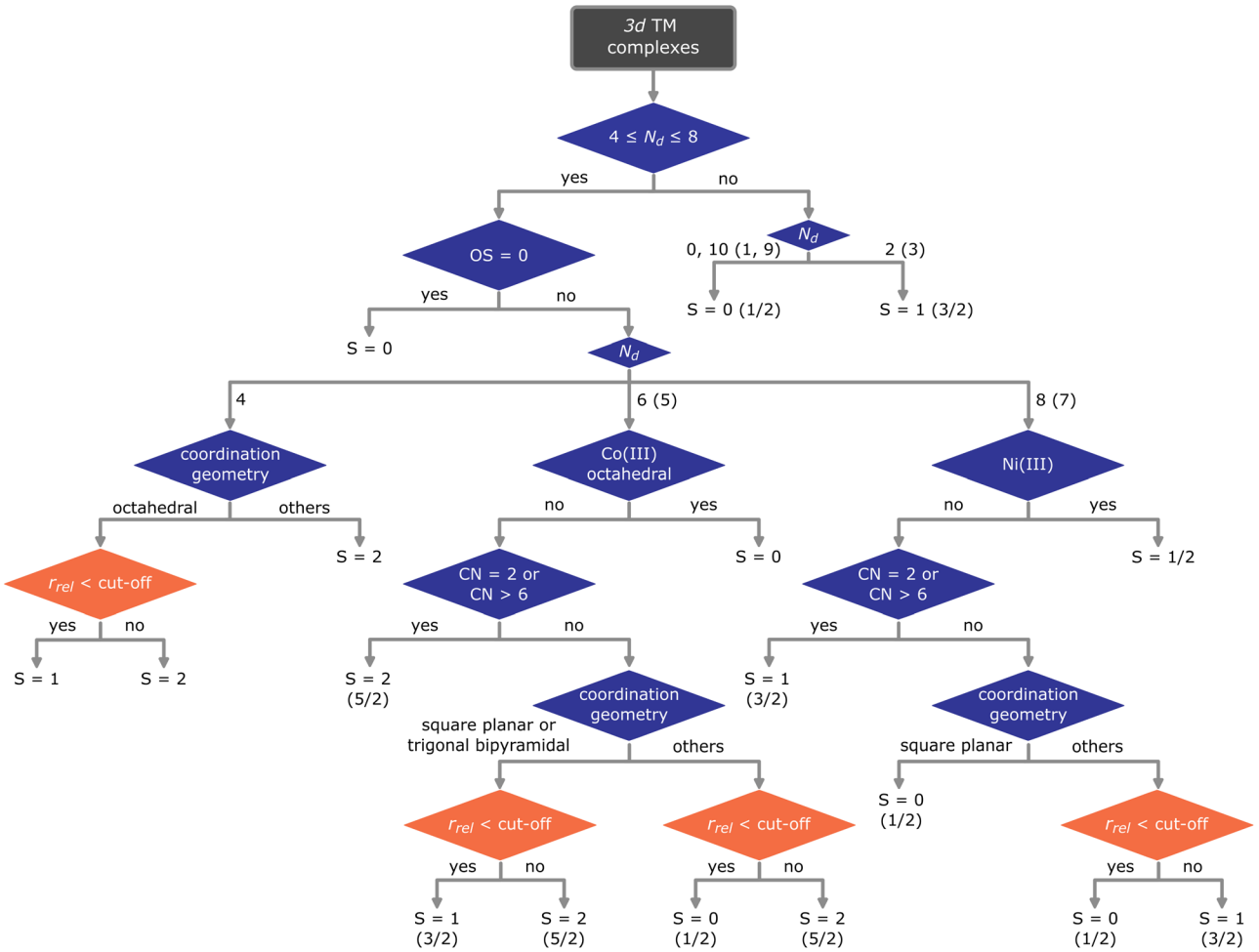

Fig. 4 Ground state spin assignment of first row TM complexes based on empirical rules. N_d: the number of d electrons, with odd numbers of d electrons shown in parentheses, OS: metal oxidation state, CN: coordination number, r_{rel}: relative metal radius, cut-off: predefined cut-off value used to distinguish the lowest spin state. Decision nodes in orange use cut-off values that are specific for a given metal and coordination geometry (see Table S10 in the ESI†).

3.3 Ground state spin prediction with statistical models

As an alternative to using simple empirical rules, we train statistical models for ground state spin prediction, using combinations of features used in the empirical models as part of the input vectors. First, we construct a feature vector F_{TM}, containing only information about the metal center: the metal atomic number, the metal OS, and the number of d electrons. The second vector F_{CE} contains only geometric features of the coordination environment: coordination number, coordination geometry, and relative metal radius. Third, F_{TM+CE} incorporates both F_{TM} and F_{CE}. For comparison, we include the atomic Spectrum of London and Axilrod–Teller–Muto potential (aSLATM) physics-based representation[57] by using a vector that concatenates one, two, and three-body terms associated with the TM as implemented in the QML package[58] with a modified grid and cutoff consistent with our previous work.[22] Finally, $F_{TM+CE+aSLATM}$ concatenates F_{TM+CE} and aSLATM. We employ random forest models trained on the TM-GSspin dataset using all feature vectors. We also train models on individual metal subsets to assess the impact of each metal separately. An overview of the performance of these models is presented in Fig. 5.

Fig. 5a shows the 10-fold cross-validated accuracy for each random forest model, with error bars representing the standard deviation across folds. Amongst the models trained on the entire TM-GSspin dataset (Fig. 5a, leftmost), the model using F_{TM+CE} achieves the highest cross-validated accuracy, reaching 98%. In contrast, the model with the F_{CE} features exhibits the poorest performance due to the lack of metal center information. F_{CE} fails to capture the intricate relationship between ground state spin and coordination geometry, which varies depending on the TM and its OS. For comparison, the model employing aSLATM outperforms the one based on F_{CE}, benefiting from the inclusion of nuclear charge information for the metal atom. Interestingly, the F_{TM+CE} features lead to a more accurate model than aSLATM despite its simplicity. This superiority is attributed to F_{TM+CE} explicitly containing electronic information, such as the number of d electrons and metal OS—critical factors closely linked to the ground state spin of the complex, which are not explicitly captured by many-body

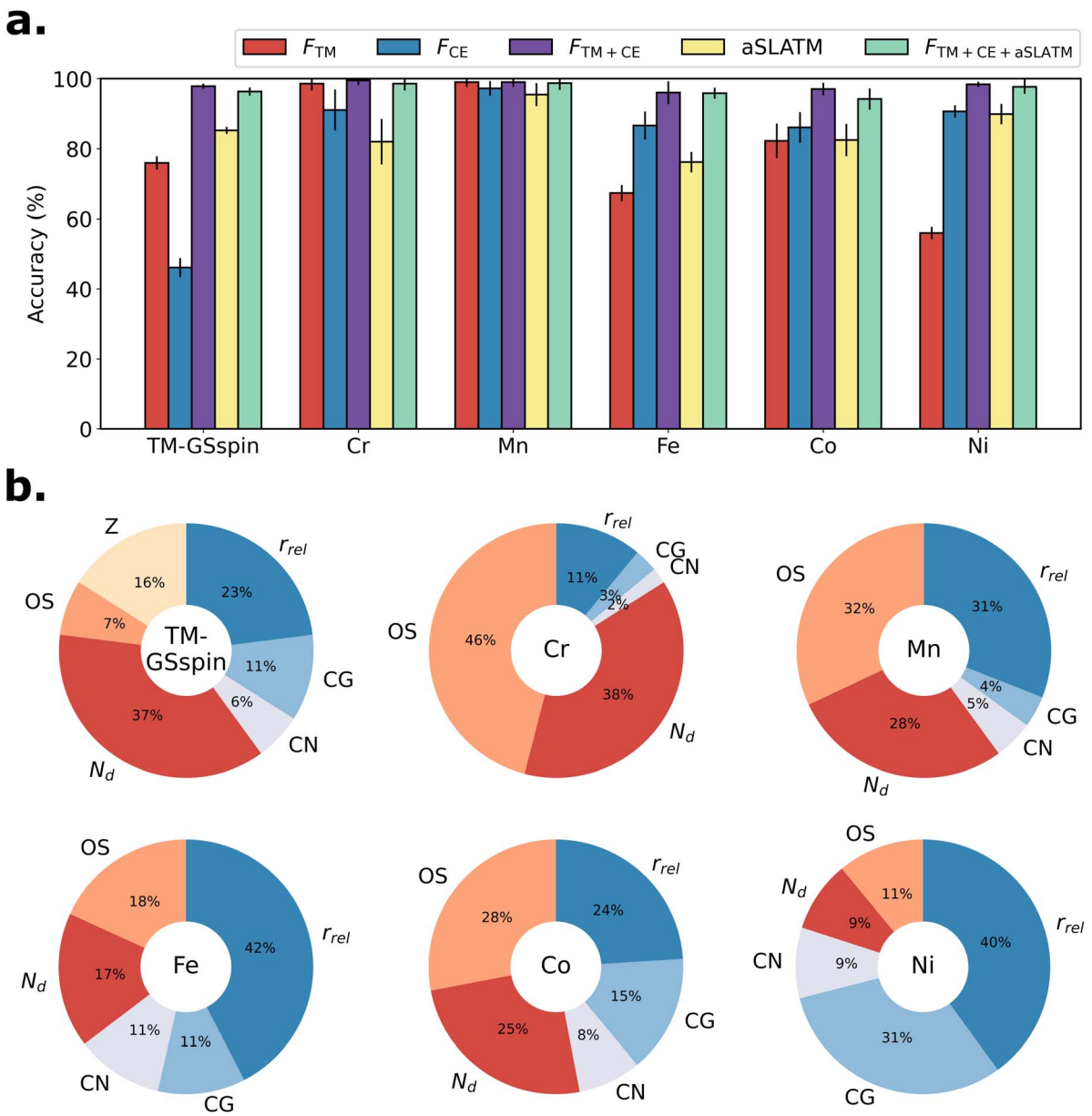

Fig. 5 Ground state spin prediction for the TM-GSspin dataset and for each metal subset. (a) 10-Fold cross-validated accuracy of random forest models using different features. F_{TM} is a vector containing the metal atomic number (Z), metal OS, and the number of d electrons (N_d). F_{CE} contains coordination number (CN), coordination geometry (CG), and relative metal radius (r_{rel}). F_{TM+CE} combines both F_{TM} and F_{CE}. aSLATM is the atomic SLATM representation[57] of the metal atom. $F_{TM+CE+aSLATM}$ incorporates F_{TM+CE} and aSLATM. (b) Feature importances in prediction models trained using F_{TM+CE}.

potential terms. When considering standard deviations, the model using F_{TM+CE} displays an accuracy similar to the model using the much larger $F_{TM+CE+aSLATM}$ (vector size 6 *vs.* 80 591) while bypassing the computational cost of generating the aSLATM representation. Overall, these results underscore the effectiveness of F_{TM+CE} in capturing both the electronic and structural information of TM complexes, crucial for determining their ground state spin. Separately, we performed dimensionality reduction on the aSLATM by using principal component analysis to reduce the number of features to 100. The reduced aSLATM resulted in slightly worse performance compared to the original aSLATM, as shown in Table S12 in the ESI.†

The performance of the models trained on the individual TM subset are also shown in Fig. 5a. The models using the F_{TM} features exhibit high accuracy for Cr and Mn complexes, owing to the strong relationship between the d electron configuration and the ground state spin for these elements. Conversely, for Fe, Co, and Ni complexes, using only F_{CE} outperforms the F_{TM} models, which is especially evident for the Ni complexes. The reason for the latter is that the majority of Ni complexes in the dataset are either Ni(II) square planar or octahedral complexes,

consistently displaying a singlet or triplet ground state, respectively. For these individual metal subsets, models using F_{CE} are comparable or even more accurate than those using aSLATM, contrasting with the trends obtained for the corresponding models trained on the entire TM-GSspin dataset. This distinction arises because the relationship between the ground state spin and the coordination environment is well-defined within each metal subset but not on the overall dataset, in which each metal exhibits different preferences. Ultimately, the best performance obtained on the individual subsets is consistently achieved for the models employing F_{TM+CE}.

To shed light on the relevance of the various features, we finally examine the feature importance derived from the F_{TM+CE} random forest models, as shown in Fig. 5b. Overall, the relative metal radius (r_{rel}) and the number of d electrons (N_d) emerge as the most influential factors. In agreement with our previous observations (*vide supra*), predictions for Cr and Mn complexes primarily rely on features associated with the metal center, while predictions for Ni complexes are driven by geometric information. Nevertheless, even in those cases, the incorporation of both electronic and structural descriptors remains crucial to predict the ground state spins of TM complexes in both individual metal subsets or full dataset. For more comprehensive information regarding the performance of random forest models, a detailed list of complexes with incorrect predictions, and the analysis of feature importances in models trained using $F_{TM+CE+aSLATM}$, see Tables S13, S14, and Fig. S18 in the ESI.†

4 Conclusion

In order to facilitate the high-throughput generation of computed data by retrieving information from existing crystallographic data repositories, we here present ground state spin prediction models that will extend the capability of our *cell2mol*[22] software. *cell2mol* was built to offer a comprehensive interpretation of the unit cell by extracting the connectivity and total charge of molecules within a crystal structure, placing emphasis on transition metal-containing structures. Yet, information regarding another necessary input of quantum chemical computations, *i.e.*, the molecular spin, was missing from the interpretation. Here, we propose a general approach to predict the ground state spin of TM complexes. The TM-GSspin dataset, comprising 2063 mononuclear first row TM complexes and their computed ground state spins, was constructed starting from the 31k complexes extracted from the CSD with *cell2mol*. TM-GSspin is currently the largest dataset that encompasses a diverse range of metals, metal OS, coordination geometries, and first coordination sphere compositions, while also containing total charge and ground state spin information.

The analysis of TM-GSspin uncovered correlations between ground state spins and the various features of TM complexes (*e.g.*, metal OS, the number of d electrons, coordination number, geometry, and the relative metal radius measure introduced herein). While most of the relationships were already established, we quantified their validity across the board and exploited them to build rule-based decision trees to assign the ground state spin for first row TM complexes. The most relevant features were also used as inputs of random forest models, achieving an impressive 98% cross-validated accuracy within the dataset.

These models are fully integrated into the latest version of *cell2mol* which is now capable of determining the total charge, the OS, and the ground state spin of TM complexes directly from crystallographic data. This work streamlines automation of electronic structure workflows of molecules extracted from crystal structure repositories.

Data availability

The TM-GSspin dataset as well as the additional complexes discussed in the ESI† are available in the Materials Cloud Repository **https://doi.org/10.24435/materialscloud:jx-a5**. The ground state spin prediction models are available in the development version of *cell2mol* **https://github.com/lcmd-epfl/cell2mol/tree/dev**. Random forest models are trained and tested by using the Python script **https://github.com/lcmd-epfl/cell2mol/blob/dev/cell2mol/random_forest.py**.

Author contributions

Y. C. and C. C. conceived the project. Y. C. curated the dataset, performed the DFT computations, constructed empirical and statistical models, and analyzed the results with help from R. L. and S. V. All authors discussed the results. The original manuscript was written by Y. C. with help and feedback from all authors. C. C. provided supervision throughout and was responsible for funding acquisition.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

This research was supported by the National Centre of Competence in Research (NCCR) MARVEL (grant number 205602), a NCCR funded by the Swiss National Science Foundation (SNSF). The NCCR Catalysis (grant number 180544) of SNSF is also acknowledged for financial support of R. L.

References

1 R. Pollice, G. dos Passos Gomes, M. Aldeghi, R. J. Hickman, M. Krenn, C. Lavigne, M. Lindner-D'Addario, A. Nigam, C. T. Ser, Z. Yao, *et al.*, *Acc. Chem. Res.*, 2021, **54**, 849–860.
2 A. Nandy, C. Duan, M. G. Taylor, F. Liu, A. H. Steeves and H. J. Kulik, *Chem. Rev.*, 2021, **121**, 9927–10000.
3 R. Gómez-Bombarelli and A. Aspuru-Guzik, *Handbook of Materials Modeling: Methods: Theory and Modeling*, 2020, pp. 1939–1962.
4 S. Curtarolo, G. L. Hart, M. B. Nardelli, N. Mingo, S. Sanvito and O. Levy, *Nat. Mater.*, 2013, **12**, 191–201.

5 J. E. Saal, S. Kirklin, M. Aykol, B. Meredig and C. Wolverton, *JOM*, 2013, **65**, 1501–1509.
6 E. O. Pyzer-Knapp, C. Suh, R. Gómez-Bombarelli, J. Aguilera-Iparraguirre and A. Aspuru-Guzik, *Annu. Rev. Mater. Res.*, 2015, **45**, 195–216.
7 R. Gómez-Bombarelli, J. Aguilera-Iparraguirre, T. D. Hirzel, D. Duvenaud, D. Maclaurin, M. A. Blood-Forsythe, H. S. Chae, M. Einzinger, D.-G. Ha, T. Wu, *et al.*, *Nat. Mater.*, 2016, **15**, 1120–1127.
8 J. T. Blaskovits, R. Laplaza, S. Vela and C. Corminboeuf, *Adv. Mater.*, 2024, **36**, 2305602.
9 C. Hölzer, I. Gordiy, S. Grimme and M. Bursch, *J. Chem. Inf. Model.*, 2024, **64**, 825–836.
10 K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev and A. Walsh, *Nature*, 2018, **559**, 547–555.
11 M. Rupp, A. Tkatchenko, K.-R. Müller and O. A. Von Lilienfeld, *Phys. Rev. Lett.*, 2012, **108**, 058301.
12 G. B. Goh, N. O. Hodas and A. Vishnu, *J. Comput. Chem.*, 2017, **38**, 1291–1307.
13 J. A. Keith, V. Vassilev-Galindo, B. Cheng, S. Chmiela, M. Gastegger, K.-R. Müller and A. Tkatchenko, *Chem. Rev.*, 2021, **121**, 9816–9872.
14 D. Morgan and R. Jacobs, *Annu. Rev. Mater. Res.*, 2020, **50**, 71–103.
15 B. Huang and O. A. Von Lilienfeld, *Chem. Rev.*, 2021, **121**, 10001–10036.
16 J. P. Janet, L. Chan and H. J. Kulik, *J. Phys. Chem. Lett.*, 2018, **9**, 1064–1071.
17 A. C. Mater and M. L. Coote, *J. Chem. Inf. Model.*, 2019, **59**, 2545–2559.
18 H. J. Kulik, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2020, **10**, e1439.
19 A. G. Garrison, J. Heras-Domingo, J. R. Kitchin, G. dos Passos Gomes, Z. W. Ulissi and S. M. Blau, *J. Chem. Inf. Model.*, 2023, **63**, 7642–7654.
20 C. Bo, F. Maseras and N. López, *Nat. Catal.*, 2018, **1**, 809–810.
21 M. Nakata and T. Shimazaki, *J. Chem. Inf. Model.*, 2017, **57**, 1300–1308.
22 S. Vela, R. Laplaza, Y. Cho and C. Corminboeuf, *npj Comput. Mater.*, 2022, **8**, 188.
23 I. Bruno, S. Gražulis, J. R. Helliwell, S. N. Kabekkodu, B. McMahon and J. Westbrook, *Data Sci. J.*, 2017, **16**, 38.
24 D. Balcells and B. B. Skjelstad, *J. Chem. Inf. Model.*, 2020, **60**, 6135–6146.
25 M. G. Taylor, T. Yang, S. Lin, A. Nandy, J. P. Janet, C. Duan and H. J. Kulik, *J. Phys. Chem. A*, 2020, **124**, 3286–3299.
26 S. Gallarati, P. van Gerwen, R. Laplaza, S. Vela, A. Fabrizio and C. Corminboeuf, *Chem. Sci.*, 2022, **13**, 13782–13794.
27 C. R. Groom and F. H. Allen, *Angew. Chem., Int. Ed.*, 2014, **53**, 662–671.
28 C. R. Groom, I. J. Bruno, M. P. Lightfoot and S. C. Ward, *Acta Crystallogr., Sect. B: Struct. Sci., Cryst. Eng. Mater.*, 2016, **72**, 171–179.
29 K. M. Jablonka, D. Ongari, S. M. Moosavi and B. Smit, *Nat. Chem.*, 2021, **13**, 771–777.
30 M. Swart, *J. Chem. Theory Comput.*, 2008, **4**, 2057–2066.
31 M. Swart and M. Gruden, *Acc. Chem. Res.*, 2016, **49**, 2690–2697.
32 J. P. Janet, C. Duan, A. Nandy, F. Liu and H. J. Kulik, *Acc. Chem. Res.*, 2021, **54**, 532–545.
33 J. P. Janet and H. J. Kulik, *Chem. Sci.*, 2017, **8**, 5137–5152.
34 E. I. Ioannidis and H. J. Kulik, *J. Chem. Phys.*, 2015, **143**, 034104.
35 A. Nandy, D. B. Chu, D. R. Harper, C. Duan, N. Arunachalam, Y. Cytter and H. J. Kulik, *Phys. Chem. Chem. Phys.*, 2020, **22**, 19326–19341.
36 M. Reiher, O. Salomon and B. Artur Hess, *Theor. Chem. Acc.*, 2001, **107**, 48–55.
37 O. Salomon, M. Reiher and B. A. Hess, *J. Chem. Phys.*, 2002, **117**, 4729–4737.
38 C. E. Housecroft and A. G. Sharpe, *Inorganic chemistry*, Pearson Education, 2008, vol. 1.
39 P. Atkins, *Shriver and Atkins' inorganic chemistry*, Oxford University Press, USA, 2010.
40 R. H. Crabtree, *The organometallic chemistry of the transition metals*, John Wiley & Sons, 2009.
41 K. P. Butin, E. K. Beloglazkina and N. V. Zyk, *Russ. Chem. Rev.*, 2005, **74**, 531.
42 E. Bernuz Fito, *PhD thesis*, Universitat de Barcelona, 2021.
43 M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. A. Montgomery Jr, J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J. M. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, Ö. Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski and D. J. Fox, *Gaussian 09 Revision D.01*, Gaussian Inc., Wallingford CT, 2009.
44 S. Grimme, J. Antony, S. Ehrlich and H. Krieg, *J. Chem. Phys.*, 2010, **132**, 154104.
45 S. Grimme, S. Ehrlich and L. Goerigk, *J. Comput. Chem.*, 2011, **32**, 1456–1465.
46 S. Vela, M. Fumanal, J. Cirera and J. Ribas-Arino, *Phys. Chem. Chem. Phys.*, 2020, **22**, 4938–4945.
47 V. Vennelakanti, M. G. Taylor, A. Nandy, C. Duan and H. J. Kulik, *J. Chem. Phys.*, 2023, **159**, 024120.
48 N. J. DeYonker, K. A. Peterson, G. Steyl, A. K. Wilson and T. R. Cundari, *J. Phys. Chem. A*, 2007, **111**, 11269–11277.
49 K. D. Vogiatzis, M. V. Polynski, J. K. Kirkland, J. Townsend, A. Hashemi, C. Liu and E. A. Pidko, *Chem. Rev.*, 2018, **119**, 2453–2523.
50 M. Radoń, *Phys. Chem. Chem. Phys.*, 2019, **21**, 4854–4870.

51 P. Gütlich, Y. Garcia and H. A. Goodwin, *Chem. Soc. Rev.*, 2000, **29**, 419–427.
52 J. A. Real, A. B. Gaspar and M. C. Muñoz, *Dalton Trans.*, 2005, 2062–2079.
53 G. Xu and X. Li, *Acta Crystallogr., Sect. E: Struct. Rep. Online*, 2010, **66**, m544.
54 C.-Y. Lin, J. C. Fettinger, N. F. Chilton, A. Formanuik, F. Grandjean, G. J. Long and P. P. Power, *Chem. Commun.*, 2015, **51**, 13275–13278.
55 B. Cordero, V. Gómez, A. E. Platero-Prats, M. Revés, J. Echeverría, E. Cremades, F. Barragán and S. Alvarez, *Dalton Trans.*, 2008, 2832–2838.
56 D. W. Blakesley, S. C. Payne and K. S. Hagen, *Inorg. Chem.*, 2000, **39**, 1979–1989.
57 B. Huang and O. A. von Lilienfeld, *Nat. Chem.*, 2020, **12**, 945–951.
58 A. S. Christensen, F. Faber, B. Huang, L. Bratholm, A. Tkatchenko, K. Muller and O. von Lilienfeld, QML: A Python toolkit for quantum machine learning, 2017, **https://github.com/qmlcode/qml**.