TMtopo Dataset — Quantum Geometries and Density Topology for 1.1k Transition Metal Complexes

Filipe Teixeira,* Edgar Silva-Santos, and M. Natália D. S. Cordeiro*

LAQV-REQUIMTE, Faculty of Sciences of the University of Porto, Rua do Campo Alegre, 4169-007 Porto, Portugal

E-mail: filipe.teixeira@fc.up.pt; ncordeir@fc.up.pt

Abstract

The chemistry of transition metal (TM) complexes is a hugely diverse field where commonly accepted chemistry concepts are routinely challenged, hindering the development of predictive models of TM complex chemistry. In this work, we report the TMtopo data set containing optimized geometries, quantum calculated properties, and quantum topological descriptors for 1110 first row TM complexes. Properties were computed at the TPSSh/Def2-TZVP level of theory and the quantum topological descriptors were collected under the framework of the Quantum Theory of Atoms in Molecules (QTAIM), including a systematic topological survey of the Laplacian of the electron density, $\nabla^2\rho(\mathbf{r})$. This survey yielded novel insights on the proliferation of inner Valence Shell Charge Concentrations (iVSCCs, local minima of $\nabla^2\rho(\mathbf{r})$) in the metal center, suggesting that their number is determinant for the stabilization of the metal center in a more intense manner than their arrangement opposing each of the metal's ligands (*Inorg. Chem.* **2016**, *55*, 3653). Pairwise representation of the collected properties revealed overall low correlation, although some structure could be perceived in

the data (specially when considering the topological features of $\nabla^2\rho(\mathbf{r})$). This suggests that the TMtopo data set could be usefully exploited in the data-driven discovery of new TM complexes with interesting properties for applications in as catalysis, opto-electronics and sustainable energy production and storage. TMtopo is an open data set that can be accessed free of charge from https://github.com/teixeirafilipe/TMtopo.

1 Introduction

Machine Learning (ML) is playing an important role in chemical research[1–6]. ML algorithms generate predictive models mapping a set of descriptors (features) into one or more properties (targets) by minimizing the error relative to reference training data[7]. In order to accomplish this, ML techniques such as Artificial Neural Networks[8,9], Kernel Ridge Regression[2], Support Vector Machines[5], and Random Forests[10,11] rely on highly flexible algorithms with a myriad of internal parameters, and thus require big data sets for their training, as well as to access their predictive capabilities regarding new data[11,12]. This poses an important hurdle when applying ML to chemical research, as reference experimental data is usually small and limited in the exploration of the chemical space. In order to circumvent this obstacle, ML models started to be trained with data generated from first-principles Quantum Mechanical (QM) calculations, giving rise to quantum-based ML (QML)[2,12–16]. QML models are used to predict energies and nuclear gradients (enabling ML-accelerated molecular dynamics simulations)[14,17,18], but also other properties such as the energies of the frontier molecular orbitals, dipole moments, polarizabilities, band structure, and other quantum properties[19,20]. Similar to their experimentally-rooted counterparts, these QML models also require large and comprehensive data sets to avoid biasing and overfitting issues. Examples of such data sets are the Materials Project[21] and PubChemQC[22], which cover significant parts of the chemical space for solid state and general (mainly main-group) chemistry, respectively. Quantum data sets devoted to transition metal (TM) complexes have been known to cover either small[23] or very specific regions[24] of the chemical space. Recently, Balcells and Skjelstad[25] published the tmQM dataset containing geometries of about 86 000 TM complexes, as well as a few QM-calculated properties, such as the total energy, HOMO and LUMO energies, Natural charge of the metal center, norm of the dipole momentum and polarizability.

One important aspect of TM chemistry is the relatively large diversity of molecular geometries attainable even by compounds with relatively simple molecular formula (i.e., low coordination number and mono-atomic ligands). Often, the most stable geometry of these

complexes does not conform to the predictions made by well established models such as VSEPR, ligand field, crystal field, or ligand-repulsion models[26]. In this regard, Gillespie and co-workers noticed that the Electron Localization Function (ELF) of several TM fluorides, oxofluorides, hydrides and methanides has some interesting characteristics, namely the absence of Electron Localization Basins (ELB) at the spatial region where the metal's valence shell is expected, as well as the appearance of ELB's in the region of the outermost core shell and located opposite to each ligand[27,28]. These results where confirmed by analysing the Laplacian of the electron density, $\nabla^2\rho(\mathbf{r})$, which shows local maxima ($(3,+3)$ critical points) in the same spacial region and geometrical arrangement[29]. These observations had lead to the conclusion that the penultimate shell (usually referred to as the inner-Valence Shell, iVS) of the metal center of a TM complex is not isotropic and that the molecular geometry in these complexes is dictated by the lowering of the repulsion among these iVS Charge Concentrations (iVSCC) that lie opposite to each ligand.[26–28]

More recently, a detailed analysis of several vanadium-acetate complexes under the more framework of the Quantum Theory of Atoms in Molecules (QTAIM) has revealed some important exceptions of Gillespie's observations[30]. Although vanadium-acetate complexes in which acetate behaved as a mono-dentate ligand usually conformed to Gillespie's observations, the ones with at least one bidentate acetate ligand did not. What is more, when partitioning the molecular energy by the individual atomic basins under the QTAIM framework, the former complexes show a lower energy for the metal center, whereas the latter show a stabilization of the atoms in the acetate moiety, due to the electron delocalization along the V−O−C−O ring[30]. These results lead to the hypothesis that QTAIM properties derived from the analysis of both the electron density, $\rho(\mathbf{r})$, as well as $\nabla^2\rho(\mathbf{r})$ could be useful for creating novel ML models predicting molecular geometries, as well as other important aspects of TM-complexes.

We herein report the TM topological database (TMtopo)[31], which contains a curated collection of first-row TM complexes containing O, F and F. The following sections describe

the computational methodology used to generate the data, the architecture of the database, as well as a statistical overview of the data contained in TMtopo. The results and data are shared in the hope of nurturing the development of novel data-driven (including ML) models for the advancement of TM complex chemistry and their application.

2 Methodology

2.1 Exploration of the Chemical Subspace

The TMtopo data comprises the equilibrium geometries of all fluoride, chloride and oxygen (oxo-) complexes of first-row TMs (Sc to Zn), with a general formula $MO_iCl_jF_k$. The chemical subspace includes all common oxidation states for each of the metals, from +1 (Cu and Sc) to +6 (Cr, Mn, Fe and Co), as well as the lowest lying states of the spin multiplicities accessible for each complex, given the formal occupation of the $3d$ shell: hexaplet for d^5; quintuplet for d^4 and d^6; quadruplet for d^3, d^5, and d^7; triplet for d^2, d^4, d^6, and d^8; doublet for d^1, d^3, d^5, d^7, and d^9; and singlet for d^0, d^2, d^4, d^6, d^8, and d^{10} metal centers.

In order to generate all possible geometries, the following procedure was implemented using an in-house developed Python script. Given the metal and the number of oxygen, fluoride and chloride atoms, the metal center was placed at the origin of the Cartesian coordinates. Then, all possible arrangements of the ligand atoms were generated by placing each ligand atom on different vertices of a virtual octahedron centerd at the origin and oriented so that the vertices lie on the xx, yy and zz axes, at a distance of 1.6 Å from the center. Following that, all redundant geometries were discarded, the metal-chloride bonds were stretched to 2.2 Å, and the structure was saved as an initial guess for the subsequent geometry optimization step. It should be noted that this strategy is unable to yield trigonal planar or trigonal bipyramidal guesses when the coordination number (n_{coord}) is 3 or 5, respectively. Because of this, a similar procedure was carried out for the TM complexes with 3 or 5 ligands, starting from an idealized triangular bipyramid, and the redundant structures

(mainly T-shaped molecules) were discarded. At the end of these procedures, 1649 guess structures were generated (considering variations in spin multiplicity).

2.2 Quantum Geometries and Properties

Each guess structure was subjected to geometry optimization using Density Functional Theory (DFT) calculations at the TPSSh approximation[32,33], given the good performance it attained in several benchmarks using TM complexes[33–36]. The Def2-TZVP basis set from Ahlrichs[37] was used for all atoms. Vibrational analysis of the equilibrium geometries allowed the selection of geometries representing true minima of the Potential Energy Surface (PES), for which all vibrational frequencies are positive. All DFT calculations were carried out using version 4.0.1.2 of the Orca program package[38]. The resulting geometries were then grouped by molecular formula and spin multiplicity, and each group was scanned for redundant equilibrium geometries, taking into account all bond lengths, all valence angles centerd on the metal atom as well as the out-of-plane angles in complexes for which $n_{coord} > 2$, taking the plane defined by the metal center and the first two ligands as reference.

After discarding all non-equilibrium and redundant structures, the resulting set of 1110 TM complexes was further analysed under the QTAIM framework, using the AIMAll software, version 16.08.17[39]. For each structure, relevant information on the topology of $\rho(\mathbf{r})$ and $\nabla^2\rho(\mathbf{r})$ was collected, as well as atomic basin populations, effective atomic charges, and energy decomposition analysis over the atomic basins.

The data was collected in structured text files (one file per compound), containing a description of the TM complex and the level of theory used, equilibrium geometry, vibrational and thermochemistry data, as well as several QTAIM data pertaining to the Bond Critical Points (BCPs) found, properties of the atomic basins, as well as information on the minima of $\nabla^2\rho(\mathbf{r})$ (points of local charge concentration). A detailed overview of the structure of these entries is given in the Supporting Information (SI), as on the Github page hosting the data base: https://github.com/teixeirafilipe/TMtopo.

2.3 Data Availability

The TMtopo database is an open data set freely available at GitHub[31], comprising DFT-optimized geometries and quantum properties calculated at the TPSSh/Def2-TZVP level of theory, as well as an assortment of quantum topological descriptors derived from the analysis of $\rho(\mathbf{r})$ and $\nabla^2\rho(\mathbf{r})$. In addition to this, database specifications and example code for handling the data is also provided in GitHub (https://github.com/teixeirafilipe/TMtopo), and in the SI.

3 Results and Discussion

The TMtopo database comprises calculated geometries and quantum properties for 1110 TM complexes with oxygen (oxo-), fluoride and chloride. The distribution of the entries in the data along the first-row transition metals depends heavily on the oxidation states available for each metal, as shown in Figure 1a. At the same time, the composition of the data also reflects the fact that metals in higher oxidation states are able to form a more diverse set of complexes (at least in theory), given the limited number of ligands considered in this work. Although some TM complexes of high spin multiplicity were discarded due to poor convergence of the self-consistent field calculations, or due to the presence of at least one negative (imaginary) vibrational frequency in their predicted infra-red spectrum, the distribution along the different spin multiplicities reflects the availability of each spin state given the formal occupation of the $3d$ shell of each metal, at each oxidation state, as shown in Figure 1b.

Figure 1c depicts the representation of the different molecular geometries in the data, and the relative proportion of each metal center within each geometry. Again, the representation of a given molecular geometry appears to be limited only by the availability of the underlying coordination number(s), which in turn depends on the distribution of the complexes by metal and oxidation state. The distribution depicted in Figure 1c suggests that the geometry

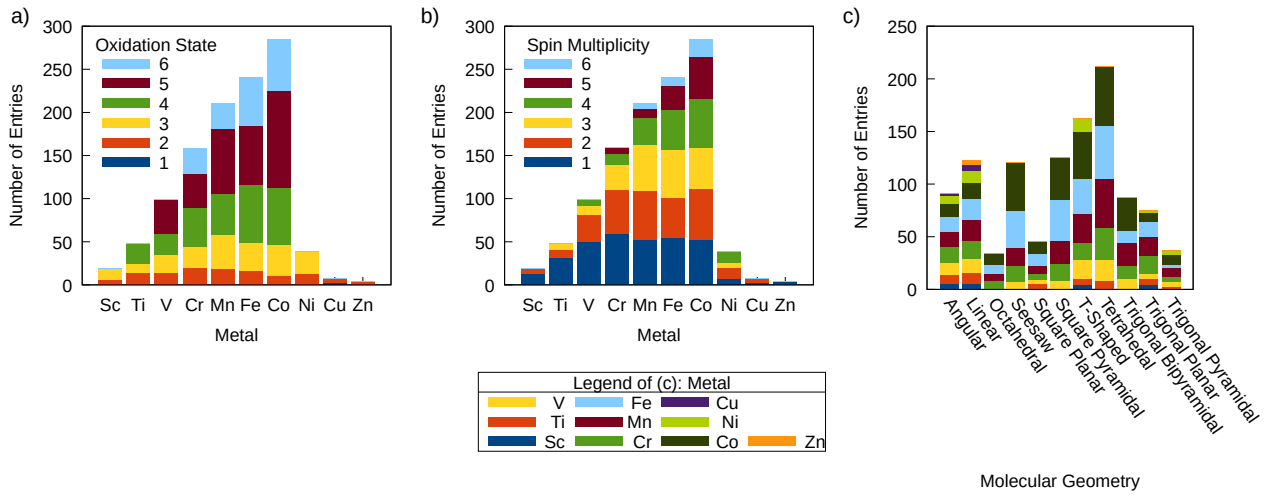

Figure 1: Distribution of the entries in the data: (a) distribution by metal center and its formal oxidation state; (b) distribution by metal center and spin multiplicity; (c) distribution by molecular geometry and by metal center. The height of the bars shows the cumulative number of entries in each of the categories shown in the horizontal axis.

optimization procedure was able to locate different minima of the PES, depending on the initial guess geometry. This is well illustrated by the relatively high frequency of T-shaped TM complexes ($n_{coord} = 3$), as well as square planar ($n_{coord} = 4$), and square pyramidal $n_{coord} = 5$ TM complexes. However, it must be stressed that the single criteria for inclusion of a given TM complex in the database is being a minima in the PES at a given spin multiplicity, and does not endorse a particular form of a TM complex as being neither the most stable one, nor synthetically available.

On the other hand, the TMtopo database allows one to evaluate and correlate quantum properties of TM complexes, in the prospect of training predictive ML models for evaluating these properties in larger or more complex compounds. The database was designed to address the possibility of training predictive models of interesting quantum properties from data concerning $\rho(\mathbf{r})$ and $\nabla^2\rho(\mathbf{r})$, and in particular the configuration and characteristics of the iVSCCs on the metal center. Thus, an overview regarding the distribution of these data is in order. For each metal, the iVS is located in a region of space between 0.25 Å and 0.45 Å from the TM nucleus. As shown in Figure 2a, the location of the iVS mainly reflects the increase in nuclear charge, with the distribution of the iVS radius (r_{iVS}) being mostly

mono-modal for each metal. An important exception to this is provided by Ti, which shows a bi-modal distribution. A closer inspection of the data for the Ti complexes highlighted some factors that may contribute to a more contracted iVS layer, such has a triplet state for Ti(II), or the presence of chloride ligands. However, the data does not allow for a clear ruling regarding these observations, and one may postulate that the lower effective nuclear charge in Ti makes the iVS more sensitive to the atom's chemical environment.

Three to eight iVSCC were found in within the QTAIM atomic basin of the metal center of each TM complex. In the case of ScO, 18 additional minima of $\nabla^2\rho(\mathbf{r})$ were found further away from the nucleus, at about 1.9 Å. These points were discarded as numerical artefacts, given the very low value of $\rho(\mathbf{r})$ in that region. As shown in Figure 2b, Sc, Ti and V mostly exhibit four or five iVSCCs, whereas Cr, Mn, Fe and Co show a stronger tendency to present six iVSCCs. In the case of the latter metals (Ni, Cu and Zn), the tendency seems to show eight iVSCCs. Overall, the data depicted in Figure 2b shows a distinct tendency to exhibit a larger number of iVSCCs when progressing along the first TM series. What is more, the number of iVSCCs does not appear to vary considerably with n_{coord}, as illustrated in Figure 2c. Indeed, Figure 2c shows a dominance of complexes with $n_{coord} = 4$, but also the preference for the number of iVSCCs (n_{iVSCC}) equal to six, irrespective of n_{coord}. This preference for $n_{\mathrm{iVSCC}} = 6$ is also observed irrespective of the geometry of the TM complex, as illustrated in Figure 2d, with the notable exception of trigonal planar complexes, which apparently prefer $n_{\mathrm{iVSCC}} = 5$.

The data collected in this work appears to contradict the observations made by Gillespie and co-workers[27–29] regarding the preference for an iVSCC configuration where each iVSCC lies opposite to a ligand. According to such observations, the spacial arrangement of the iVSCC should match that of the TM geometry. As illustrated in Figure 3, this is only observed for a small portion of the TM complexes in the TMtopo database. Indeed, the iVSCCs can achieve a large variety of spacial arrangements, which do not appear to be clearly related to the geometry of the TM complex. Nevertheless, it is important to notice

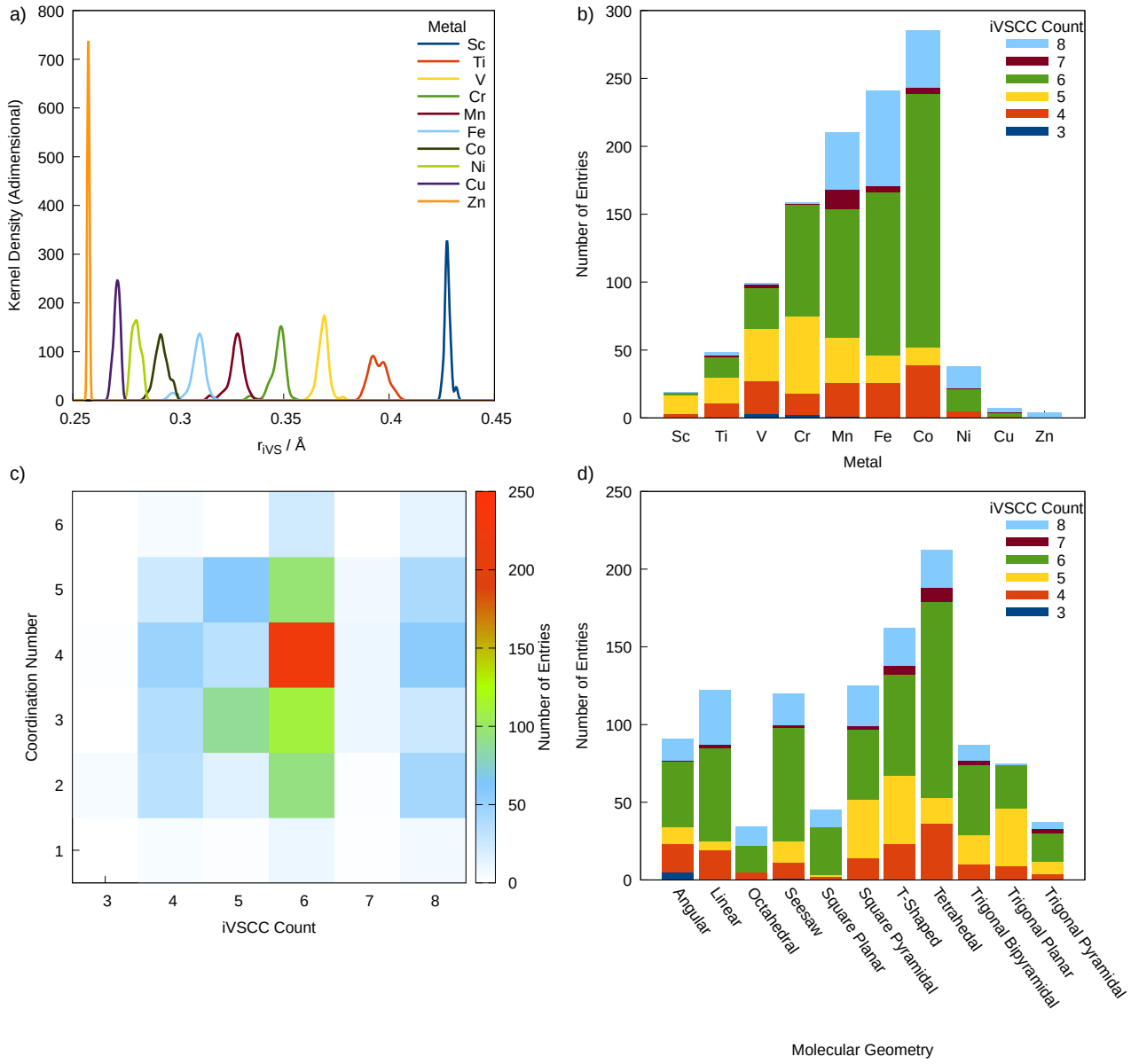

Figure 2: Distribution of the data regarding the characteristics of the metal's inner Valence Shell (iVS) in TMtopo: (a) kernel density distribution of the iVS radius (r_{iVS}) by metal; (b) number of iVSCCs (n_{iVSCC}) by metal center; (c) cross-distribution of n_{iVSCC} and n_{coord}; (d) distribution of n_{iVSCC} along the different observed molecular geometries.

that tetrahedral arrangements of the iVSCCs are usually found in tetrahedral TM complexes. What is more, square pyramidal complexes are quite likely to also have a square pyramidal arrangement of the iVSCCs. Trigonal bipyramidal complexes also appear to follow Gillespie's observations and prefer a trigonal bipyramidal arrangement of the iVSCCs.

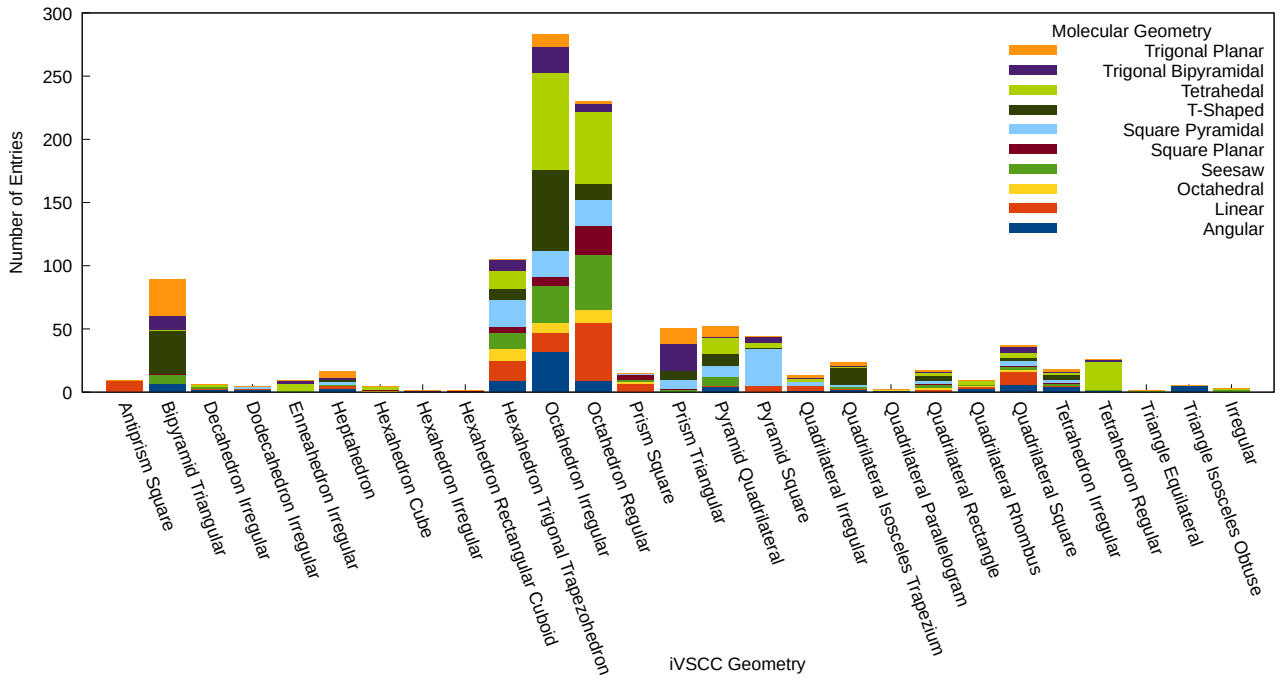

Figure 3: Distribution of the data regarding the spacial of the iVSCC and the geometry of the TM-complexes in the TMtopo database.

Further observation of the topology of $\nabla^2\rho(\mathbf{r})$ has shown that at least some iVSCCs could be found in a location opposite to a ligand, even if not all iVSCCs in a particular complexes follows that rule. This is unavoidable when $n_{\text{iVSCC}} > n_{coord}$, which accounts for a large portion of such cases. More rarely, one would notice that the arrangement of the iVSCCs follows Gillespie's rule for some but not all ligands, even if $n_{\text{iVSCC}} = n_{coord}$. These observations raised the need to quantify how much a given TM complex deviates from Gillespie's rule. For this purpose, each iVSCC was classified as being gillespian if and only if a Metal–Ligand BCP was found along the line that connects the iVSCC to the metal nucleus (with a tolerance of 10° to accommodate eventual numerical noise when searching for the critical points of $\nabla^2\rho(\mathbf{r})$). This allows the definition of two index measuring adherence to

Gillespie's rule:

$$G_1 = \frac{n_g}{n_{\mathrm{iVSCC}}} \tag{1}$$

and,

$$G_2 = \frac{n_g}{n_{coord}} \tag{2}$$

where n_g is the number of gillespian iVSCCs found in the complex. Figure 4 depicts the relative abundance of each $(G_1, G_2$ combination in the data, showing that most species in the database do not follow Gillespie's observations at all (i.e. $G_1 = 0.0$, and $G_2 = 0.0$, accounting for 695 entries), and only 54 compounds are completely gillespian (i.e. $G_1 = G_2 = 1.0$). More interesting, 361 TM-metals have intermediary values of either indexes. Within this latter group, two tendencies are perceptible in Figure 4: some compounds do appear to have extra iVSCCs than was is expected from the TM-geometry ($G_1 \leq 0.0$, but $G_2 = 1.0$), while other complexes appear to follow Gillespie's observations only with respect to some of its ligands. Also worth noting, is the absence of cases where ($G_1 = 1.0$, but $G_2 < 1.0$), which would correspond to all iVSCCs being opposed to a ligand, but not all ligands being meet with an opposing iVSCC. Such an absence leads to the conclusion that the cases where arrangement of the iVSCCs follows Gillespie's rule for some but not all ligands (mentioned above) is never observed due to a lack of iVSCCs in the metal basin.

These observations might reflect the tendency for attaining certain values of n_{iVSCC}, as shown in Figures 2c and 2d, which overcomes the preference for a gillespian arrangement of the iVSCCs. In order evaluate the importance of this trend, Figure 5a depicts the distribution of the total molecular energy per electron according to n_{iVSCC}. Despite the large overlapping between the different populations divided by their n_{iVSCC}, Figure 5a strongly suggests that complexes with $n_{\mathrm{iVSCC}} = 5$ might be less stable than an isoelectronic complex with a different arrangement and count of iVSCCs.

The relatively large dispersion of $E_{\mathrm{Mol}}/n_{\mathrm{electrons}}$ shown in Figure 5a is justifiable by the relatively small impact the iVSCC should have on the potential felt by the electrons in the

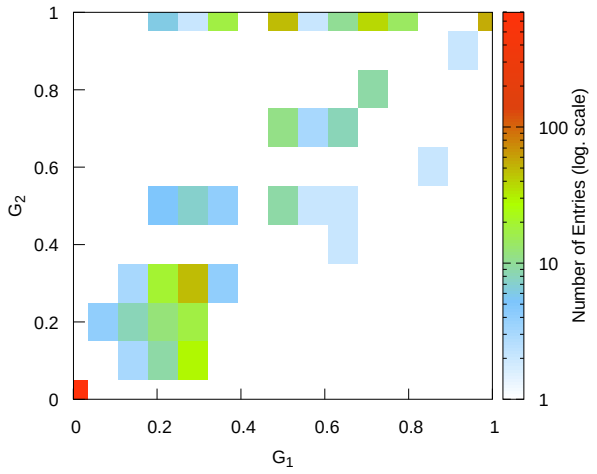

Figure 4: Distribution of the G_1 and G_2 in the TMtopo database.

atomic basins pertaining to the ligands of each TM complex. Although relying on a local adherence to the virial theorem, the negative value of the electronic kinetic energy in the metal basin, $-K(\mathrm{M})$, is a suitable surrogate for contribution of the metal's population to the electronic energy of the molecule, specially when considering general trends in the data[30,40]. Figure 5b depicts $-K(\mathrm{M})$ per electron in the metal's basin, which highlights the impact of n_{iVSCC} on the stability of the metal center (and then to the contribution of the metal center to the total energy). The energy distribution per electron shown in Figure 5b clearly suggest that odd values of n_{iVSCC} are linked to higher electronic energy of the metal center, with the case where $n_{\mathrm{iVSCC}} = 3$ being particularly penalizing. On the other hand, even values of n_{iVSCC} are linked to lower energy per electron, especially when $n_{\mathrm{iVSCC}} = 8$ and $n_{\mathrm{iVSCC}} = 6$. Hence, the data shows that the strong presence of these latter n_{iVSCC} values shown in Figure 2c reflects an underlying physical cause, possibly a lowering in the inter-electronic repulsion.

What is more, Figure 5b further suggests that the energetic penalty associated with an odd n_{iVSCC} can range from 5 to 10 E_h per electron. This value is two to three orders of magnitude higher than the energetic penalty observed for non-gillespian arrangements of the iVSCC in vanadium-acetate complexes[30], suggesting a precedence of n_{iVSCC} over the geometrical arrangement of the coordination sphere.

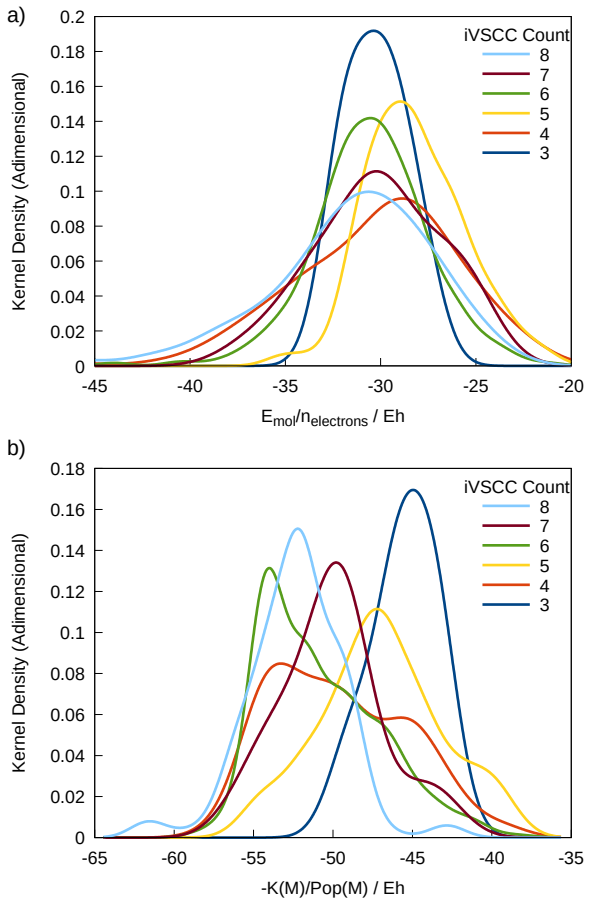

Figure 5: Kernel density estimates of the distribution of the total electronic energy per electron ($E_{\mathrm{Mol}}/n_{\mathrm{electrons}}$) (a), and of the distribution of the electronic kinetic energy per electron among the electrons within the metal's atomic basin ($-K(\mathrm{M})/Pop(\mathrm{M})$) (b). In both cases the data is subdivided by n_{iVSCC}.

Given our main objective of providing a database of quantum calculated properties and topological features of $\rho(\mathbf{r})$ and $\nabla^2\rho(\mathbf{r})$ for a large number of first-row TM complexes, the remainder of this discussion will explore the nature of the data by representing commonly used quantum property pairs in scatter plots. Figure 6a shows that the pairwise distribution of the HOMO and LUMO energies has some structure, with some Fe, Co, Ni and Zn complexes showing higher LUMO energies, yielding also large HOMO-LUMO gaps. These abnormalities are also reflected in Figure 6b, which displays the polarizability versus the HOMO-LUMO gap. Nonetheless, the correlation between the two properties appears to be low. Each point in 6b is colored by the value of the G_1, highlighting a concentration of complexes with high G_1 in the region corresponding to a HOMO-LUMO gap between 3.0 and 6.0 eV.

In general, there is poor correlation between the QTAIM charge of the metal center, q_{metal}, and the magnitude of the dipole moment ($|\mu_{\text{Dipole}}|$), as displayed in Figure 6c. Indeed, the spread of the scatter plot displayed in Figure 6c closely resembles the scattering of the same properties in the tmQM dataset, recently published by Balcells and Skjelstad[25]. One should notice, however, that the highest scores of the G_1 index, are concentrated amongst the higher values of q_{metal}, providing some structure to these data. In a similar fashion, q_{metal} also bears very low correlation with the isotropic polarizability, as shown in Figure 6d. Moreover, Figure 6b further suggests poor correlation between the G_2 index and either properties.

Figure 7 further explores the relationship between G_1, G_2 and some quantum calculated properties of interest: HOMO-LUMO gap, $|\mu_{\text{Dipole}}|$, and the isotropic polarizability and quadrupole. Overall, the representations displayed in Figure 7 demonstrate low correlation between these four important properties and either G_1 or G_2, but do suggest some structure in the data, which might be worthy of further exploration. For example, Figure 7a suggests a region in the HOMO-LUMO gap between 4.0 and 6.0 eV that is mainly occupied by compounds with high G_1 and $G_2 = 1.0$. On the other hand, Figure 7b shows a tendency for

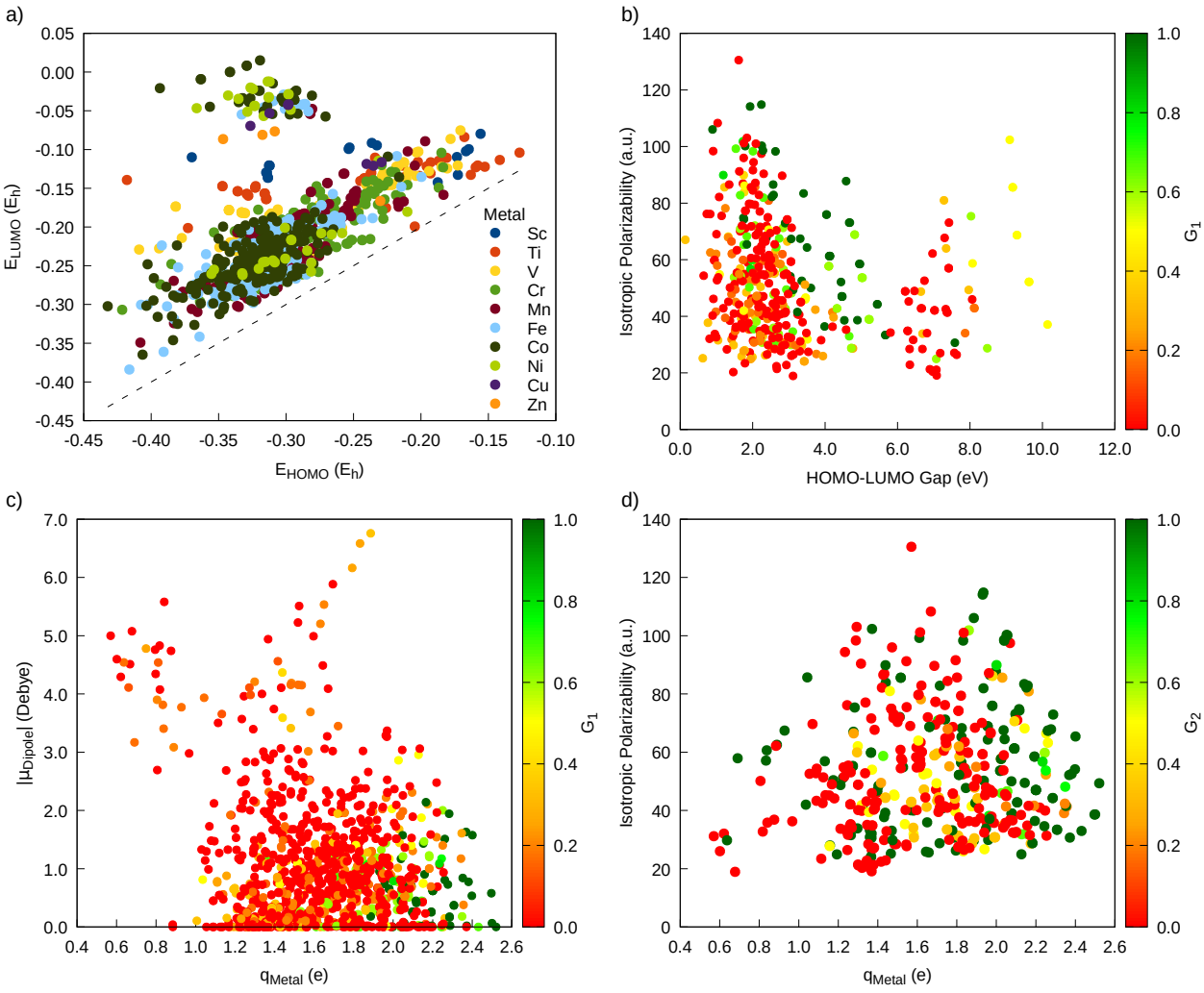

Figure 6: Pairwise representation of commonly used quantum properties: a) HOMO and LUMO energies, colored by metal; b) HOMO-LUMO gap *versus* the isotropic polarizability, colored by the value of the G_1 index; c) QTAIM charge of the metal (q_{metal}) *versus* the magnitude of the dipole moment ($|\mu_{\text{Dipole}}|$), colored by G_1, and; d) q_{metal} *versus* the isotropic polarizability, colored by the value of the G_2 index.

gillespian compounds to bear lower dipole moments, which may be due to a more symmetrical arrangement of both the ligands and the iVSCCs. Figure 7c shows some ascending trend of the isotropic polarizability with respect to G_1, although strongly obfuscated by other factors.

Furthermore, Figure 7d reveals some interesting structure concerning the data for the isotropic quadrupole: although both gillespian and completely non-gillespian (i.e. $G_1 = G_2 = 0$) compounds appear to fill the range available for this variable, TM complexes with $G_2 = 1$ and intermediate values of G_1 prefer higher (less negative) values. TM complexes with values of G_1, but $G_2 = 1$ also show higher values of the isotropic quadrupole, but lowering G_2 does increase the dispersion to a region almost as wide as the one observed for $G_1 = G_2 = 0$. Because of this, the graphic displayed in Figure 7d shows a void region, suggesting that compounds with $G_2 = 1$ do not present an isotropic quadrupole lower than -50 a.u., unless G_1 is also 1.

In general, the G_1 and G_2 indexes convey some information on the properties of TM complexes in the TMtopo database. The structured data suggested in Figure 7 is not easily discernible when considering other iVS descriptors, such as n_{iVSCC} and r_{iVS}, as depicted in Figure S1 of the SI. Indeed, r_{iVS} is clearly related to the atomic number of the metal center (Cf. Figure 2a), whereas n_{iVSCC} appears to follow a general trend towards eight iVSCCs arranged in an octahedral arrangement, as discussed above.

4 Conclusions

This work presents the TMtopo data set, which provides geometries, quantum calculated quantities and quantum topological descriptors for over 1000 TM complexes containing one first-row transition metal center and any possible combination of fluoride, chloride and oxygen, in the form of the oxo ligand. The complexes were systematically generated in order to explore different possible molecular geometries and spin multiplicities, and optimized using well established DFT techniques. The data was then curated in order to exclude duplicated

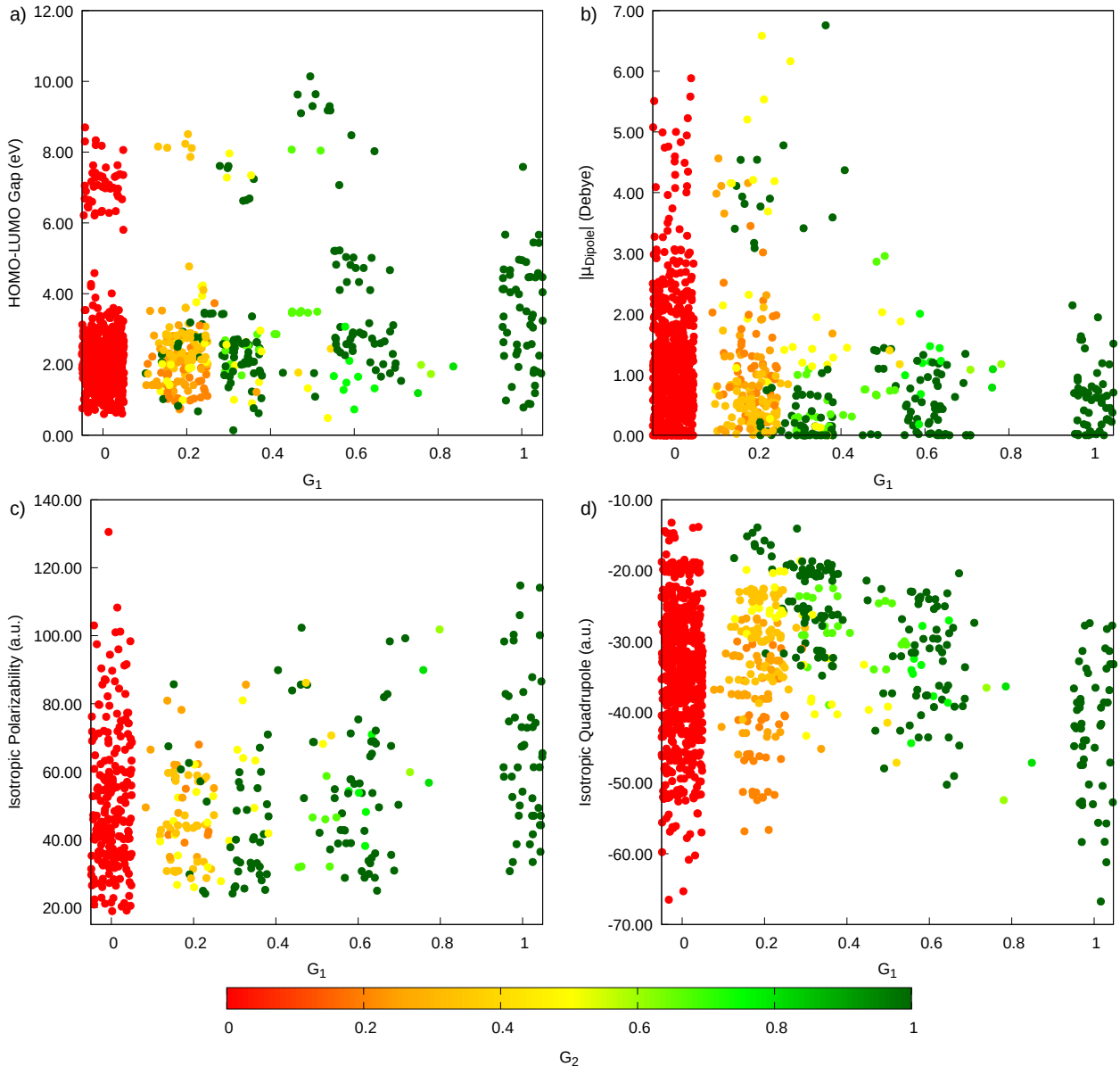

Figure 7: Scatter plots of the HOMO-LUMO gap (a); $|\mu_{\text{Dipole}}|$ (b); isotropic polarizability (c), and isotropic quadrupole (d), against the G_1 index, and colored by the value of G_2 (common color code at the bottom of the Figure). Jitter with an amplitude of 0.1 were introduced along the xx coordinates (G_1, adimensional) in order to aid visualization.

and non-equilibrium geometries. A total of 1110 unique entries were compiled upon curation. Quantum topological descriptors were then collected under the QTAIM framework.

Particular attention was given to the topological features of $\nabla^2\rho(\mathbf{r})$ within the metal's atomic basin, which allowed for an unprecedented systematic survey of the properties concerning the relationship between iVSCC and other molecular properties. Preliminary observations reported in this work highlight that the number of iVSCCs is of paramount importance in lowering the electronic energy of the metal center, allowing for non-gillespian arrangements of the iVSCCs to proliferate. Deviations to the gillespian arrangements were quantified in the form of two indexes G_1 and G_2, measuring excess of iVSCCs and their misalignment with the ligands, respectively.

Pairwise representations of the data in TMtopo suggests a large number of poorly correlated descriptor/property, although some structure is noticeable in such representations, suggesting that further useful information may be gathered by advanced multivariate statistics and/or machine learning methods. We hope to report on the application of these methods for gathering further information in a subsequent publication.

Supporting Information Available

Additional data (database specifications and additional figures) are available free of charge in the Electronic Supporting Information. This information is available free of charge via the Internet at http://pubs.acs.org.

Acknowledgement

This work received financial support from PT national funds (FCT/MCTES, Fundação para a Ciência e Tecnologia and Ministsério da Ciência, Tecnologia e Ensino Superior) through the project UIDB/50006/2020, as well as through research project REALM — Reactive Learning Machines PTDC/QUI-QIN/30649/2017. ES further thanks FCT/MCTES for a

325 scholarship under project PTDC/ART-OUT/29692/2017.

References

(1) Gromski, P. S.; Henson, A. B.; Granda, J. M.; Cronin, L. How to explore chemical space using algorithms and automation. *Nat. Rev. Chem.* **2019**, *3*, 119–128.

(2) Fabrizio, A.; Meyer, B.; Fabregat, R.; Corminboeuf, C. Quantum Chemistry Meets Machine Learning. *CHIMIA Int. J. Chem.* **2019**, *73*, 983–989.

(3) Yang, W.; Fidelis, T. T.; Sun, W.-H. Machine Learning in Catalysis, From Proposal to Practicing. *ACS Omega* **2019**, *5*, 83–88.

(4) Kitchin, J. R. Machine learning in catalysis. *Nat. Catal.* **2018**, *1*, 230–232.

(5) Chen, C.; Zuo, Y.; Ye, W.; Li, X.; Deng, Z.; Ong, S. P. A Critical Review of Machine Learning of Energy Materials. *Adv. Energy Mater.* **2020**, *10*, 1903242.

(6) Orupattur, N. V.; Mushrif, S. H.; Prasad, V. Catalytic materials and chemistry development using a synergistic combination of machine learning and ab initio methods. *Comput. Mater. Sci.* **2020**, *174*, 109474.

(7) Coley, C. W.; Barzilay, R.; Jaakkola, T. S.; Green, W. H.; Jensen, K. F. Prediction of Organic Reaction Outcomes Using Machine Learning. *ACS Cen. Sci.* **2017**, *3*, 434–443.

(8) Li, H.; Zhang, Z.; Liu, Z. Application of Artificial Neural Networks for Catalysis: A Review. *Catalysts* **2017**, *7*, 306.

(9) Amabilino, S.; Bratholm, L. A.; Bennie, S. J.; Vaucher, A. C.; Reiher, M.; Glowacki, D. R. Training Neural Nets To Learn Reactive Potential Energy Surfaces Using Interactive Quantum Chemistry in Virtual Reality. *J. Phys. Chem. A* **2019**, *123*, 4486–4499.

(10) Strobl, C.; Boulesteix, A.-L.; Kneib, T.; Augustin, T.; Zeileis, A. Conditional Variable Importance for Random Forests. *BMC Bioinformatics* **2008**, *9*.

(11) Breiman, L. *Mach. Learn.* **2001**, *45*, 5–32.

(12) Toyao, T.; Maeno, Z.; Takakusagi, S.; Kamachi, T.; Takigawa, I.; ichi Shimizu, K. Machine Learning for Catalysis Informatics: Recent Applications and Prospects. *ACS Catal.* **2019**, *10*, 2260–2297.

(13) McDonagh, J. L.; Silva, A. F.; Vincent, M. A.; Popelier, P. L. A. Machine Learning of Dynamic Electron Correlation Energies from Topological Atoms. *J. Chem. Theory Comput.* **2017**, *14*, 216–224.

(14) Zielinski, F.; Maxwell, P. I.; Fletcher, T. L.; Davie, S. J.; Pasquale, N. D.; Cardamone, S.; Mills, M. J. L.; Popelier, P. L. A. Geometry Optimization with Machine Trained Topological Atoms. *Sci. Rep.* **2017**, *7*, 12817.

(15) Rupp, M. Machine learning for quantum mechanics in a nutshell. *Int. J. Quantum Chem.* **2015**, *115*, 1058–1073.

(16) Hansen, K.; Montavon, G.; Biegler, F.; Fazli, S.; Rupp, M.; Scheffler, M.; von Lilienfeld, O. A.; Tkatchenko, A.; Müller, K.-R. Assessment and Validation of Machine Learning Methods for Predicting Molecular Atomization Energies. *J. Chem. Theory Comput.* **2013**, *9*, 3404–3419.

(17) Brockherde, F.; Vogt, L.; Li, L.; Tuckerman, M. E.; Burke, K.; Müller, K.-R. Bypassing the Kohn-Sham equations with machine learning. *Nat. Commun.* **2017**, *8*, 872.

(18) Behler, J. Perspective: Machine learning potentials for atomistic simulations. *J. Chem. Phys.* **2016**, *145*, 170901.

(19) Faber, F. A.; Hutchison, L.; Huang, B.; Gilmer, J.; Schoenholz, S. S.; Dahl, G. E.; Vinyals, O.; Kearnes, S.; Riley, P. F.; von Lilienfeld, O. A. Prediction Errors of Molecu-

lar Machine Learning Models Lower than Hybrid DFT Error. *J. Chem. Theory Comput.* **2017**, *13*, 5255–5264.

(20) Taylor, M. G.; Yang, T.; Lin, S.; Nandy, A.; Janet, J. P.; Duan, C.; Kulik, H. J. Seeing Is Believing: Experimental Spin States from Machine Learning Model Structure Predictions. *J. Phys. Chem. A* **2020**, *124*, 3286–3299.

(21) Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; Persson, K. A. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Mater.* **2013**, *1*, 011002.

(22) Nakata, M.; Shimazaki, T. PubChemQC Project: A Large-Scale First-Principles Electronic Structure Database for Data-Driven Chemistry. *J. Chem. Inf. Model.* **2017**, *57*, 1300–1308.

(23) Friederich, P.; dos Passos Gomes, G.; De Bin, R.; Aspuru-Guzik, A.; Balcells, D. Machine learning dihydrogen activation in the chemical space surrounding Vaska's complex. *Chem. Sci.* **2020**, *11*, 4584–4601.

(24) Janet, J. P.; Ramesh, S.; Duan, C.; Kulik, H. J. Accurate Multiobjective Design in a Space of Millions of Transition Metal Complexes with Neural-Network-Driven Efficient Global Optimization. *ACS Cent. Sci.* **2020**, *6*, 513–524.

(25) Balcells, D.; Skjelstad, B. B. tmQM Dataset—Quantum Geometries and Properties of 86k Transition Metal Complexes. *J. Chem. Inf. Model.* **2020**, *60*, 6135–6146.

(26) Gillespie, R. J. Improving our understanding of molecular geometry and the {VSEPR} model through the ligand close-packing model and the analysis of electron density distributions. *Coord. Chem. Rev.* **2000**, *197*, 51 – 69.

(27) Bytheway, I.; Gillespie, R. J.; Tang, T.-H.; Bader, R. F. W. Core Distortions and Geometries of the Difluorides and Dihydrides of Ca, Sr, and Ba. *Inorg. Chem.* **1995**, *34*, 2407–2414.

(28) Gillespie, R. J.; Bytheway, I.; Tang, T.-H.; Bader, R. F. W. Geometry of the Fluorides, Oxofluorides, Hydrides, and Methanides of Vanadium(V), Chromium(VI), and Molybdenum(VI): Understanding the Geometry of Non-VSEPR Molecules in Terms of Core Distortion. *Inorg. Chem.* **1996**, *35*, 3954–3963, PMID: 11666589.

(29) Gillespie, R. J.; Robinson, E. A. Models of molecular geometry. *Chem. Soc. Rev.* **2005**, *34*, 396–407.

(30) Teixeira, F.; Mosquera, R.; Melo, A.; Freire, C.; Cordeiro, M. N. D. S. Roots of Acetate-Vanadium Linkage Isomerism: A QTAIM Study. *Inorg. Chem.* **2016**, *55*, 3653–3662.

(31) Teixeira, F. TMtopo - Properties and Quantum Topological Descriptors for Transition Metal Complexes. 2017; `https://github.com/teixeirafilipe/TMtopo`.

(32) Tao, J.; Perdew, J. P.; Staroverov, V. N.; Scuseria, G. E. Climbing the Density Functional Ladder: Nonempirical Meta-Generalized Gradient Approximation Designed for Molecules and Solids. *Phys. Rev. Lett.* **2003**, *91*, 146401–146405.

(33) Staroverov, V. N.; Scuseria, G. E.; Tao, J.; Perdew, J. P. Comparative assessment of a new nonempirical density functional: Molecules and hydrogen-bonded complexes. *J. Chem. Phys.* **2003**, *119*, 12129–12137.

(34) Minenkov, Y.; Singstad, A.; Occhipinti, G.; Jensen, V. R. The accuracy of DFT-optimized geometries of functional transition metal compounds: a validation study of catalysts for olefin metathesis and other reactions in the homogeneous phase. *Dalton Trans.* **2012**, *41*, 5526–5541.

(35) Jensen, K. P. Bioinorganic Chemistry Modeled with the TPSSh Density Functional. *Inorg. Chem.* **2008**, *47*, 10357–10365.

(36) Kossmann, S.; Kirchner, B.; Neese, F. Performance of modern density functional theory for the prediction of hyperfine structure: meta-GGA and double hybrid functionals. *Mol. Phys.* **2007**, *105*, 2049–2071.

(37) Weigend, F.; Ahlrichs, R. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297–3305.

(38) Neese, F. Software update: the ORCA program system, version 4.0. *WIREs Comput. Mol. Sci.* **2017**, *8*, e1327.

(39) Keith, T. A. AIMAll (Version 16.08.17). 2016; `http://aim.tkgristmill.com/`.

(40) Teixeira, F.; Mosquera, R.; Melo, A.; Freire, C.; Cordeiro, M. N. D. S. Driving Forces in the Sharpless Epoxidation Reaction: A Coupled AIMD/QTAIM Study. *Inorg. Chem.* **2017**, *56*, 2124–2134.

Graphical TOC Entry