

RESEARCH ARTICLE OPEN



cell2mol: encoding chemistry to interpret crystallographic data

Sergi Vela^{1,2}, Ruben Laplaza^{1,3}, Yuri Cho^{1,2} and Clémence Corminboeuf^{1,2,3}✉

The creation and maintenance of crystallographic data repositories is one of the greatest data-related achievements in chemistry. Platforms such as the Cambridge Structural Database host what is likely the most diverse collection of synthesizable molecules. If properly mined, they could be the basis for the large-scale exploration of new regions of the chemical space using quantum chemistry (QC). Yet, it is currently challenging to retrieve all the necessary information for QC codes based exclusively on the available structural data, especially for transition metal complexes. To overcome this limitation, we present *cell2mol*, a software that interprets crystallographic data and retrieves the connectivity and total charge of molecules, including the oxidation state (OS) of metal atoms. We demonstrate that *cell2mol* outperforms other popular methods at assigning the metal OS, while offering a comprehensive interpretation of the unit cell. The code is made available, as well as reliable QC-ready databases totaling 31k transition metal complexes and 13k ligands that contain incomparable chemical diversity.

npj Computational Materials (2022)8:188; <https://doi.org/10.1038/s41524-022-00874-9>

INTRODUCTION

Datasets have become an essential part of computational chemistry^{1–11}. Improvements in data availability, interoperability, and storage have been key to the development of cost-efficient Machine Learning (ML) methods¹², and promoted the use of Quantum Chemistry (QC) computations for the high-throughput screening of molecules and materials. For datasets to be adequate for both ML and QC, they must cover a well-defined albeit diverse portion of the chemical space, and include the essential information needed to run an electronic structure computation, namely the structure (*R*), and the molecular charge (*Q*). Unfortunately, satisfying both requisites can be very challenging. Existing crystallographic databases (e.g., CSD^{13,14}, COD¹⁵) offer the most comprehensive collection of crystals resulting from decades of creativity in synthetic chemistry work. They represent an ideal basis for the data-driven discovery of materials, but in these *top-down* databases, there is no information about *Q* for individual molecules. For this reason, a popular alternative has been the construction of combinatorial *bottom-up* databases, in which molecules are assembled from a pool of building blocks for which *R* and *Q* are known beforehand^{16–22}. While these databases are easy to construct and can grow considerably large, they suffer from a lack of diversity in the pool of building blocks and/or in the rules to combine them. Therefore, neither approach is currently able to tackle the diversity nor information requirements to enable the use of ML and QC methods to explore somewhere close to the full chemical space.

To improve on this situation, herein we report the *cell2mol* software (available in [github](#)), a fully automatic pipeline to characterize molecular crystals that enables the construction of QC-ready datasets with large chemical diversity. The algorithm encodes chemical concepts and rules applied by chemists when interpreting crystallographic data. While *cell2mol* excels at characterizing purely-organic crystals, it is particularly useful to characterize crystals with TM complexes, which pose a bigger challenge due to their structural complexity and the multiple

oxidation states (OS) of the metal ions. The elucidation of metal OS from structural data is an active topic of research. One of the most widespread methods is the Bond-Valence Sum (BVS), which uses a set of metal- and OS-dependent parameters to capture the correlation between OS and structure²³. The BVS method can be very accurate for some metals with available parameters²⁴, but has severe limitations in its applicability (as discussed below). Recently, ML alternatives to determine the metal OS in metal-organic frameworks²⁵ and Oxygen-Coordinated Metal Atoms²⁶ have been developed. An advantage of both the reported ML models and BVS is that they are solely based on the local environment of the metal center. As such, they scale very well with the number of metal centers in a molecule, and they do not suffer when structures have experimental uncertainties (e.g., disorder, missing H atoms) as long as these are far away from a metal center. However, while knowing the metal OS (Q_M) is valuable for analysis and filtering purposes, the charge of the ligand(s) (Q_L) is still unknown and hence these methods do not provide the total charge of the complex (Q_{TMC}) and, thus, are insufficient to set up a QC computation. With this aim in mind, Parsons et al. combined the BVS method for Q_M with three other approaches that retrieve Q_L ²⁴. While the overall approach is reported to be very reliable, it still suffers from the poor applicability of BVS and requires QC computations, which defeats the purpose of characterizing Q_{TMC} .

Other approaches exist to extract more comprehensive information about crystallographic data^{27–30}. On one hand, CSD editors interpret and curate the database entries using algorithms that assess not only Q_M but also the molecular connectivity (*C*, i.e., the bond order network) and the formal atomic charges (q_i)²⁷. The goal of such algorithms is to interpret the whole unit cell, rather than only the metal center, which makes them much more valuable (e.g., for substructural or similarity searches³¹), but also more vulnerable to experimental uncertainties. Consequently, CSD algorithms have a lower—but still exceptional—ratio of success (ca. 75%) at interpreting unit cells²⁷. However, they are closed, the final *C* and Q_M data are only accessible through the limited set of export options in the CCDC software, and have only been applied

¹Laboratory for Computational Molecular Design (LCMD), Institute of Chemical Sciences and Engineering (ISIC), École Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland. ²National Centre for Computational Design and Discovery of Novel Materials (NCCR MARVEL), Ecole Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland. ³National Centre for Competence in Research “Sustainable chemical processes through catalysis (NCCR Catalysis)”, École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland. ✉email: clemence.corminboeuf@epfl.ch

to few CSD entries, so their degree of automatization is unclear. On the other hand, a fully automatized topological analysis of crystals is available with the ToposPro package³⁰ at the TopCryst web interface²⁹, which includes a simplified form of **C** based on geometry considerations. It is worth mentioning that both the CSD algorithms and ML models exploit statistics to make their predictions. In other words, the accuracy over a particular structure (or metal center) depends on how often it has been seen before, which implies that less frequent structures, metal environments, and oxidation states are less accurately described²⁵.

As opposed to these methods, *cell2mol* is able to provide a reliable and comprehensive interpretation of a unit cell, including the connectivity (**C**) and charge (**Q**) of all molecular species contained in crystallography files in a deterministic manner, meaning that it can be applied on individual.cif files with no preliminary training, as opposed to probabilistic models (i.e., ML-based²⁵). Additionally, it does not require any QC computation^{32–36}, and offers complete data interoperability³⁷. *cell2mol* provides the necessary information to set up any subsequent QC computation (including solid-state ones) or ML model, or to classify compounds/ligands based on charge, denticity (κ) -or hapticity (η)- or through any substructural search, thus retaining full control over the species included in the final datasets. To validate its performance, we construct, analyze and distribute eight different databases of TM complexes, as well as a database of 13k unique ligands with all the necessary information to exploit them in molecular assemblers to achieve even greater chemical diversity.

RESULTS AND DISCUSSION

The algorithm

The characterization of a unit cell with *cell2mol* only requires its crystallography file (i.e., the cif). After an initial formatting with the cif2cell code³⁸, the characterization proceeds in two main steps (see detailed workflow in Supplementary Note 1). The goal of step ① is to obtain the information about the stoichiometry of the different molecules in the unit cell. This information is not immediately available from the.cif file since the molecules are not yet recognized as such. Moreover, they tend to be severely fragmented in smaller groups of connected atoms (box **A** in Fig. 1). Those are put together through the construction and block-diagonalization of the adjacency matrix (**A**), with $A_{ij} = 1$, and $A_{ji} = 1$ if the distance between atoms i and j is below a threshold, otherwise zero. Thus, **A** is evaluated based on interatomic distances due to its simplicity and efficiency³⁹, while more accurate and expensive alternatives exist based on Voronoi partitioning schemes^{40,41}. After block diagonalization, the resulting blocks correspond to either molecules or fragments (i.e., a portion of a molecule). Molecules are preserved, while fragments undergo translations in the three crystal directions forming bigger fragments until all molecules are fully reconstructed (box **B**). Finally, ligands in TM complexes are identified using a similar block-diagonalization process, with $A_{Mj} = A_{jM} = 0$ for all metal (M) and ligand atoms (j) (box **C**).

In step ②, *cell2mol* proceeds to assign the connectivity (**C**) and formal charge (**Q**) for all species in the unit cell exploiting the charge neutrality rule. All ligands and any non-metallic species (i.e., counterion or lattice-solvent) are interpreted and their **C** and the associated formal total charge (**Q**) and atomic charges (q_i) are retrieved. Within *cell2mol*, the concept of connectivity is mathematically handled as the Bond-order matrix **C** representing a Lewis structure. Basically, **C** adds the bond order (e.g., single, double bond) information to **A**. The creation of **C** from **A** is done through a modified and expanded version of the xyz2mol code developed by Jensen and coworkers based on previous work by Kim⁴². The use of **C** to define the molecular graph offers improved

capabilities with respect to other codes based on **A**^{29,30}. Lewis structures give access to q_i and **Q**, and enable advanced substructural searches including specific bond patterns or functional groups. While the adopted machinery offers a clear advantage, the number of possible **C** (i.e., Lewis structures) grows very fast with increasing the number of atoms, especially in conjugated backbones. For this reason, together with the difficulty to handle periodic connectivities, *cell2mol* cannot be applied to periodic structures, for which other approaches based on **A** are available²⁹. However, the generation of **C** requires a known **Q**, which is precisely our unknown variable. Therefore, our approach is to generate several **C** starting from a list of candidate initial charges, and to select those that are plausible (box **D**). The criteria to generate and select the best **C** is the same for non-complex molecules (e.g., solvent, counterions) and ligands. However, for ligands, there is a key preliminary step. To compensate for the missing M-L bonds when generating **C**, some connected ligand atoms (i.e., those that are coordinated to the metal) must be saturated, typically with H atoms. This is a complicated part of the algorithm due to the ligand's large chemical diversity and different coordination modes. A comprehensive list of rules is specified in *cell2mol* to this task, dealing with a large variety of denticity -and hapticity- modes for both terminal and bridge ligands. For some ligands, especially polydentate ones with large, conjugated moieties, the decision to saturate connected atoms is particularly difficult, because a chemically meaningful **C** can be achieved with and without any additional protonation. Thus, multiple protonation states are created and, for each of them, several **C** are generated (see Supplementary Note 2).

From the pool of **C**, those that minimize the total number of atomic charges, and the absolute total charge before and after correction (for the removal of the added H⁺), are considered plausible and are pre-selected. Once plausible **C** are collected for all non-metal species in the unit cell, these are combined with a list of common OS for all metal species (see Supplementary Note 3), generating charge distributions. When only one charge distribution fulfils charge neutrality, it is selected, and the unit cell is successfully interpreted (box **E**). When multiple do, the unit cell is considered “unresolved” (see Supplementary Note 4 for a detailed analysis of errors in step ②). This is increasingly common in unit cells with multiple redox-active species, such as in bi- or poly-metallic complexes (A^0/B^{+1} vs. A^{+1}/B^0). Options are currently being explored to improve the interpretation of those systems (see Supplementary Note 5).

Example: YOXXUS

To illustrate the interpretation capabilities of *cell2mol*, we take YOXXUS⁴³ as an example (see Fig. 2). According to *cell2mol*, YOXXUS has four identical mono-metallic Re complexes and no counterion or solvent molecules in the unit cell. Each complex has three types of ligands. The first ligand is interpreted as being connected to the Re ion through two groups of atoms. One group consists of a substituted Cp ring with η^5 hapticity and the other is the P atom of a diphenylphosphine, with κ^1 denticity. *cell2mol* assigns this ligand a total -1 charge, after creating one protonation state, generating its connectivity under five possible charges (0, ± 1 , ± 2), and selecting the most plausible one. The second ligand is an iodine atom with -1 charge and appears twice, and the third is a neutral CO ligand, with a -1 and a $+1$ formal charge in the C and O atoms, respectively, and a triple bond between them. All this information is stored in variables and saved in a python object containing the interpretation of the whole unit cell. From this file, a user can easily export the Cartesian coordinates of the full reconstructed unit cell, as well as the cell parameters, to prepare a solid-state QC computation. Alternatively, the user can extract the **Q**, **R**, and **C** of any of the individual molecules, or that of the isolated ligands/metals/atoms

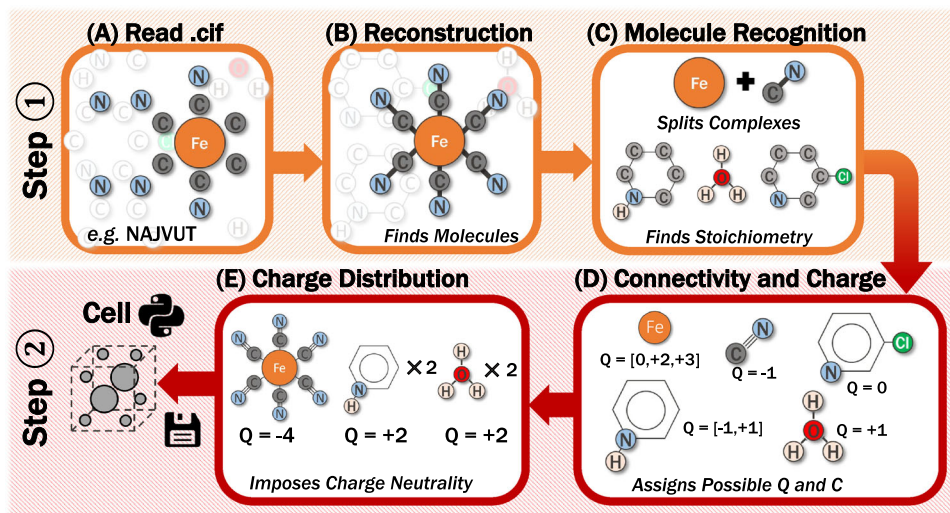


Fig. 1 Simplified workflow of *cell2mol*. Simplified workflow of *cell2mol*.

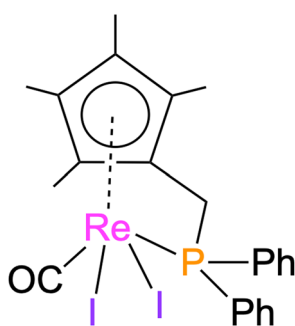


Fig. 2 Diagram of the CSD entry YOXXUS. Diagram of the CSD entry YOXXUS.

of those molecules. Indeed, non-metal species can also be accessed through their respective Rdkit *mol* objects, which provides an unprecedented level of control in the final dataset, with all the potentially relevant information (i) for a substructure/similarity search (using *C*), (ii) to set any QC computation (using *Q* and *R*), or (iii) for the generation of cheminformatics (e.g., SMILES⁴⁴) or QML-based⁴⁵ representations for ML models (using *Q*, *R*, and *C*).

Performance of *cell2mol*

The capabilities of *cell2mol* are demonstrated by interpreting crystallographic information extracted from the CSD repository. For simplicity, datasets are constructed separately for eight TM ions, including the most electronically challenging ones from the 3d block (Cr, Mn, Fe, Co, Ni, Cu) and representatives from the 4d (Ru) and 5d (Re) blocks. The data is initially extracted from the CSD software ConQuest. The only filters applied at this stage are the presence of the respective TM ion, and the absence of any so-called polymeric bond. Thus, periodic systems are discarded, for which other approaches offer excellent topological analysis tools, or the prediction of metal OS^{25,26,29}. Overall, our databases cover molecular crystals of organometallic and coordination complexes. No other limitation on the element types (except f-block), molecular size, or complexity is set. The resulting entries are exported from ConQuest in .cif format, and duplicate CSD-refcodes are discarded. Aiming at a complete interpretation of the unit cell, *cell2mol* is vulnerable to experimental uncertainties. Entries with disorder or missing H atoms cannot be interpreted correctly and are thus filtered out (see Fig. 3 and Supplementary

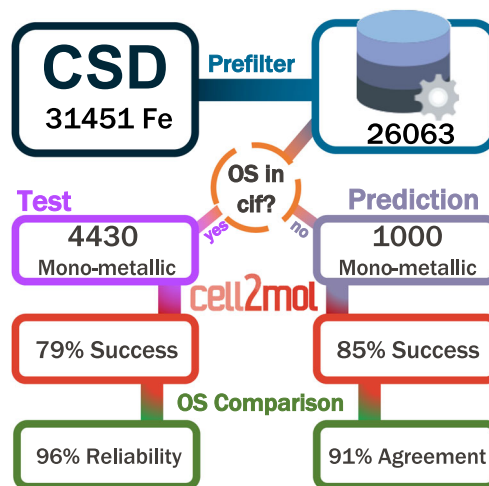


Fig. 3 General workflow of *cell2mol* performance analysis. General workflow to set up the success rate and reliability of *cell2mol*, including values for the Fe-based database of mono-metallic complexes.

Note 3). This pre-filtering step is crucial to obtain more reliable statistics of *cell2mol*. Less evident errors can only be identified after retrieving the connectivity, which is still unknown at this stage. For instance, assessing whether an O atom is missing a proton (OH vs. O vs. O⁻) depends on the connectivity (–OH vs. =O vs. –O⁻) of all molecules, and on fulfilling the charge neutrality criterion.

To evaluate the performance of *cell2mol* we use two metrics: the *success rate* and the *reliability*. The former quantifies the percentage of CSD entries for which a plausible interpretation is given, and is related to the amount of chemical diversity that the code can handle without errors. The latter, which is the most important parameter to generate curated databases, measures how often is the proposed interpretation correct. While assessing the reliability based on the entire list of properties that are extracted for each CSD entry is not possible, most have a direct impact on the assignment of the metal OS. The metal OS is thus chosen to estimate the success rate and reliability of the *cell2mol* interpretation (see Fig. 3), and is compared to the metal OS given in the .cif file, which is taken as a reference. As discussed hereafter, the reference values are sometimes erroneous, which means that

Table 1. Results on the *cell2mol* characterization of unit cells with mono-metallic TM complexes included in the test set (See Supplementary Note 5 for unit cells with bi- and poly-metallic complexes).

Metal	Total	① Success	② Success	Overall	Reliability	BVS Success	BVS Reliability	ML Reliability
Cr	1404	94%	81%	76%	94%	40%	94%	89%
Mn	2721	88%	84%	73%	96%	82%	96%	96%
Fe	4430	89%	88%	79%	96%	76%	75%	89%
Co	6554	86%	85%	73%	96%	84%	94%	94%
Ni	8478	90%	87%	78%	97%	79%	85%	96%
Cu	12811	89%	88%	78%	98%	99%	92%	96%
Ru	3522	95%	89%	84%	94%	25%	12%	94%
Re	1891	94%	75%	71%	90%	4%	42%	92%
Total	41811	91%	85%	77%	95%	61%	74%	94%
Organic	9181	97%	97%	94%	-	-	-	-

Datasets are organized by metal. The results show the *success rate* (i.e., the ratio of cases for which a method can be successfully applied) of each step of *cell2mol*, as well as the overall. For entries that result from Step ②, the metal oxidation state (OS) is compared to the one reported in the .cif file. When both values agree, we consider that the *cell2mol* interpretation is reliable. For comparison, the success for a random pool of purely-organic crystals from CSD is also shown, as well as results of the application of the BVS model (see Supplementary Note 6), and the *reliability* of the Machine-Learning (ML) model (*success* is 100%) (see Methods).

the reported reliability estimates are slightly underestimated (~1%) for all methods. Also, cases of error compensation are possible, in which the *cell2mol* interpretation is incorrect in any of its variables, while not affecting the metal OS prediction. In any case, all CSD entries for which the OS reference is available are collected in the test set, while the other entries are collected in what is called the prediction set, given that *cell2mol* predicts its properties based on the available crystallographic data. Finally, both subsets are further split in mono-, bi-, or poly-metallic complexes⁴⁶. To simplify the discussion, we focus on mono-metallic complexes although complementary analysis on the other subsets are available in Supplementary Note 5.

More than 75% of crystal structures containing mono-metallic complexes are univocally interpreted by *cell2mol* (see Table 1). This percentage raises to 94% for a pool of randomly selected purely organic crystals, and decreases to 71% for Re-based complexes, owing to their greater diversity in OSs and metal-ligand coordination modes. Such a success rate is comparable to what has been reported for the CSD interpreters, and largely outperforms other popular methods to assign the metal OS such as BVS, especially for Cr, Ru, and Re (see Supplementary Note 6). Even more important, the reliability is extraordinarily high for all metals, especially for those with one dominant OS such as Cu (98%) or Ni (97%), and diminishes to 90% in Re complexes, owing to its larger number of common OS. Entries with a disagreement are discarded from the final published datasets. However, manual inspection reveals that only about one-third of those cases are due to an error in *cell2mol* (see Supplementary Note 7). In most cases, the disagreement is due to incomplete or erroneous information in the .cif file, which suggests the potential use of *cell2mol* as a diagnostic tool. Also, the reliability of *cell2mol* is much larger than BVS (ca. 74%), which greatly underperforms here in comparison with what is typically reported in the literature, due to the much greater diversity of our datasets. Finally, to assess the performance of ML models for the same dataset, we trained a Random Forest (RF) ML model to predict the metal OS based on its local environment (see Methods for details). The accuracy of this model reaches ca. 94%, similar to what Smit and coworkers report for the application of their ML model to metal complexes (ca. 90%)²⁵, and similar to *cell2mol* itself (see Table 1). We thus conclude that *cell2mol* offers comparable reliability, while providing not only the metal OS (such as the ML and BVS methods) but a comprehensive interpretation of the unit cell. The advantages of *cell2mol* become

even clearer when interpreting the CSD entries in the prediction set (vide infra).

Chemical diversity

For 31019 CSD entries included in the test set, *cell2mol* provided a unit cell interpretation that coincided with the metal OS provided in the .cif file. For those entries, two-dimensional maps of their chemical space have been constructed (see Methods), highlighting the charge and connectivity distribution (see Fig. 4 for Fe and Supplementary Notes 10–13 for other metals). These maps help identify structure–property correlations without any a priori assumption. For instance, (i) most Fe-based haptic compounds have Fe(II) or Fe(0) metal ions, (ii) Mn shows a clear correlation between structure and OS, or (iii) Cu complexes with coordination number 3 are almost exclusively associated with Cu(I). Overall, the eight metal centers can be found in 2407 different coordination sphere types (e.g., FeN₄O₂), and are coordinated to a pool of 13,819 unique (i.e., non-repeated) ligands with total charges that range from –6 to +2, and including 8 different hapticity modes (see Fig. 5). Those ligands are collected in a separate database that includes their coordinates, list of connected atoms, charge (Q), and bond network (C) representing their Lewis structure. On one hand, C enables us to determine, through a structural search, that in this pool of ligands there are, for instance, 6909 secondary amines, or 988 rings containing an O atom. On the other hand, the remaining data could be used to re-assemble⁴⁶ those ligands into new complexes to create a *bottom-up* database encompassing an even broader region of chemical space. For instance, the 4942 unique bi-dentate ligands in the database can be combined to generate about 20 billion octahedral complexes⁴⁸. Similarly, the re-assembled molecules could be combined, in a modular fashion⁴⁷, with the identified 1246 unique non-complex molecules (i.e., solvent, counterions) to generate new candidate unit cells fulfilling charge neutrality. While the stability and shape of these new unit cells would have to be assessed^{48,49}, we expect that *cell2mol* could also be exploited to generate chemical diversity at the supramolecular level.

Mining the CSD

We have proved that *cell2mol* is able to interpret molecular crystals and construct databases with great chemical diversity. So far, this has been done exclusively for what we defined as the test set, which amounts for ca. 50% of the total CSD entries. Ideally,

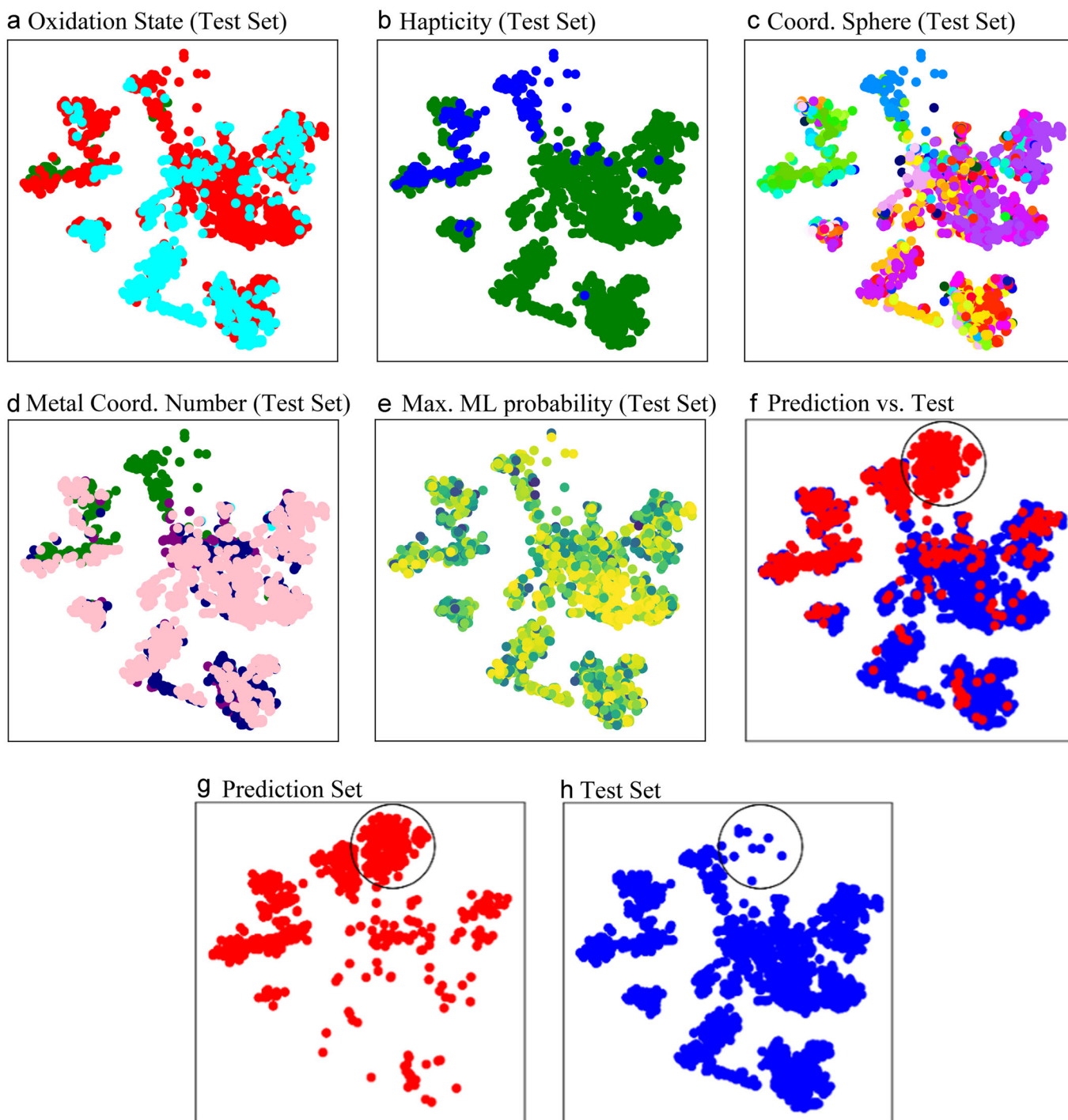


Fig. 4 Analysis of the chemical space covered by the Fe database and ML model performance. Representation of the chemical space in the Fe mono-metallic dataset using the t-SNE projection. Each point is one TM-complex in the database. Complexes are clustered by similarity in the local SLATM representation of their metal center describing the structure and chemical composition of the first coordination sphere (see Methods). In the top panel, for the test set we show **a** the distribution of metal OS (0 = green, 2 = red, 3 = cyan), Oxidation State (Test Set), **b** the presence of at least one haptic ligand (green = no, blue = yes), Hapticity (Test Set), **c** the 385 coordination sphere types for Fe, Coord. Sphere (Test Set), and **d** the coordination number of the metal, with haptic ligands counting 0 towards this number (green = 0, cyan = 3, purple = 4, navy = 5, pink = 6), Metal Coord. Number (Test Set). In the below panel, we show the **e** the maximum probability associated with the ML prediction of the metal OS, as a measure of its confidence (yellow = 1, degrading to green = 0.5 and blue = 0) Max. ML probability (Test Set), and the **f** overlap between the prediction (red) and test (blue) sets, which are also shown separately in **g** and **h**. The black circle indicates a region with poor overlap. See Supplementary Notes 10–13 for other metals.

databases would be constructed from the whole of CSD, and not be restricted to a fraction of its chemical space. Not being a statistical method, the *success rate* and *reliability* established above for the test set should, in principle, hold for the prediction set,

provided that the set of chemical rules in *cell2mol* is transferable enough. To prove it, we used 1000 mono-metallic randomly-selected CSD entries for each metal. As expected, similar *success rates* are obtained for all metals (ca. 73%, see Supplementary Note

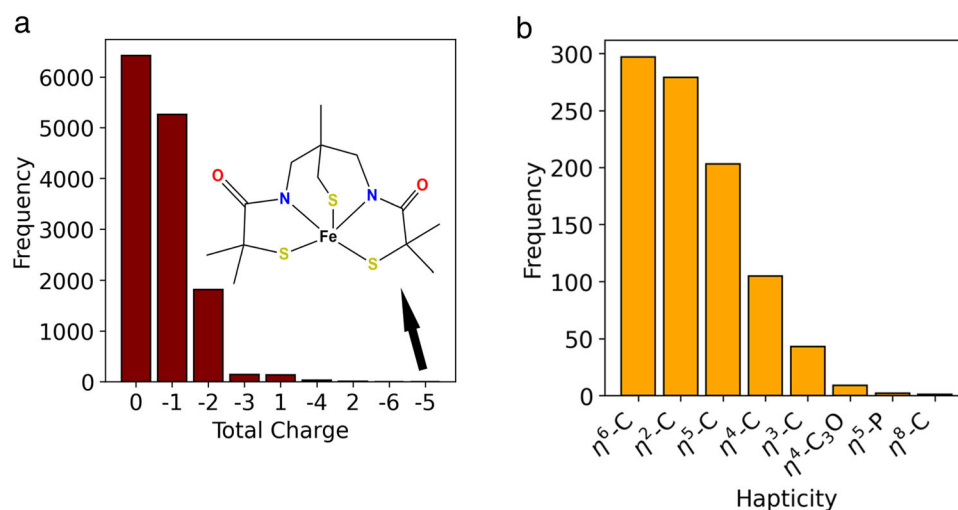


Fig. 5 Chemical diversity in the ligands database. Distribution of (left) total charges and (right) hapticity modes in the database of 13,819 unique ligands. The inset shows the single case of a ligand with -5 charge, in QORFAG.

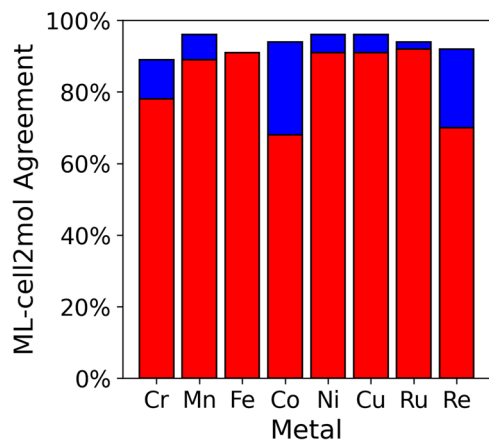


Fig. 6 Performance of the ML model for the test and prediction sets. Comparison of the performance of the trained ML model (see Methods) at predicting the metal OS. In blue, the reliability of the model, established by comparison with the.cif files in the test set. In red, the agreement between *cell2mol* and the ML model in the assignment of metal OS for the prediction set.

8). To evaluate the *reliability* in the absence of metal OS information in the.cif file, here we compared the metal OS predicted by *cell2mol* with the one provided by the ML model. Both methods coincide with about 90% of CSD entries with Mn, Fe, Ni, Cu, and Ru, which is close to the *reliability* reported for both methods in the test set (see Table 1). However, the agreement surprisingly drops to 70% for Cr, Co, and Re (see Fig. 6). Manual inspection of up to 100 cases with disagreement reveals that *cell2mol* is typically correct, which hints at deficiencies of the ML model (see Supplementary Note 9) that can be explained by the following two reasons. First, some metals exhibit a very poor correlation between structure (including chemical composition) and their OS (see Fig. 4a and Supplementary Note 10), which decreases the confidence of the ML model when assigning the OS (see Fig. 4e and Supplementary Note 12) and hence its accuracy (89% vs. 96% of agreement in Fe vs. Mn). Second, the chemical landscape can be very different in the test vs. prediction sets (see Fig. 4f–h and Supplementary Note 13), which means that the ML model often has to extrapolate. When both problems cooperate, such as in Cr, Co, and Re, the ML models lose accuracy. Considering that we used the whole available data in CSD to train this model, this behavior is likely unavoidable, and points to a

fundamental problem that statistical methods have when mining the rich chemical diversity in the entire CSD. This stresses the relevance of non-statistical alternatives such as *cell2mol*. Indeed, the most promising route for future work is the combination of a deterministic method for the comprehensive interpretation of the unit cell (e.g., *cell2mol*) with a local statistical method for the evaluation of specific properties of species when more than one possible interpretation is possible (e.g., the metal OS in *unresolved* CSD entries). Future work will focus on the implementation of this scheme, as well as on the improvement/extension of the chemical rules to understand M-L connectivity, and the incorporation of f-block metals.

Summary

We presented *cell2mol*, a tool that encodes chemical concepts and rules to interpret crystallographic data, and extract comprehensive information about the individual molecules contained in unit cells. *cell2mol* can successfully interpret about 75% of the CSD entries containing mono-metallic complexes with a reliability of over 95%. We demonstrated that these metrics surpass other popular methods dedicated to the assignment of metal OS (BVS and ML), with *cell2mol* being much more versatile. Also, we showed that our software can generate *top-down* and *bottom-up* QC-ready databases with incomparable chemical diversity. To demonstrate its capabilities, we have used *cell2mol* to generate a publicly available database of 31,019 complexes containing eight different metal centers (Cr, Mn, Fe, Co, Ni, Cu, Ru, Re). Additionally, we generated a separate database of 13,819 constituent ligands that can be rearranged to generate billions of realistic new chemical structures. All content is fully searchable and interoperable using cheminformatics software (e.g., Rdkit, SMILES-based tools). We expect that *cell2mol*, with possible subsequent improvements, will pave the way towards making all crystallographic repositories entirely usable for molecular and materials design purposes.

METHODS

CSD entries have been exported with the software ConQuest (version 5.42) included in the CCDC software, with the database updated to May 2021. The pre-filtering has been done with local bash scripts. The Random Forest (RF) model for metal-specific oxidation state prediction was constructed using the RandomForestClassifier implementation of scikit-learn⁵⁰. The local SLATM (aSLATM) representation⁵¹ of the metal center under scrutiny, aimed at capturing the structure and composition of the first

metal coordination sphere, was used as the input feature vector. All SLATM vectors were computed using the QML package⁵² with a modified discretization grid (0.4 a.u.) and cutoff (4.0 a.u.), to accommodate the requirements associated with the large number of element types. The training sets for the RF model were composed of all mono-metallic metal complexes for which a reference oxidation state was available in the CSD entry. Due to the large number of cases for Fe, Co, Ni, and Cu, these were truncated to 4000 random samples. Among the respective training sets, 10 stratified K-folds were performed for cross-validation, from which the overall out-of-sample accuracy was computed, as well as the maximum probabilities for each metal center (i.e., the probability of the assigned OS). The same local SLATM (aSLATM) representation was used to generate the t-SNE projections in Fig. 4, for which we used a perplexity value of 50.

DATA AVAILABILITY

Databases can be found at the Materials Cloud Repository (<https://doi.org/10.24435/materialscloud:g5-5r>), and the software *cell2mol* can be found in github at <https://github.com/lcmd-epfl/cell2mol>.

Received: 26 April 2022; Accepted: 10 August 2022;
Published online: 31 August 2022

REFERENCES

- Bo, C., Maseras, F. & López, N. The role of computational results databases in accelerating the discovery of catalysts. *Nat. Catal.* **1**, 809–810 (2018).
- Nandy, A., Duan, C. & Kulik, H. J. Audacity of huge: overcoming challenges of data scarcity and data quality for machine learning in computational materials discovery. *Curr. Opin. Chem. Eng.* **36**, 100778 (2022).
- Huang, B. & von Lilienfeld, O. A. Ab initio machine learning in chemical compound space. *Chem. Rev.* **121**, 10001–10036 (2021).
- Draxl, C. & Scheffler, M. NOMAD: the FAIR concept for big data-driven materials science. *MRS Bull.* **43**, 676–682 (2018).
- Smith, J. S., Isayev, O. & Roitberg, A. E. ANI-1, A data set of 20 million calculated off-equilibrium conformations for organic molecules. *Sci. Data* **4**, 170193 (2017).
- Ramakrishnan, R., Dral, P. O., Rupp, M. & von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Sci. Data* **1**, 140022 (2014).
- Jain, A. et al. Commentary: The Materials Project: a materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 11002 (2013).
- Ruddigkeit, L., van Deursen, R., Blum, L. C. & Reymond, J.-L. Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. *J. Chem. Inform. Model.* **52**, 2864–2875 (2012).
- Blum, L. C. & Reymond, J.-L. 970 Million druglike small molecules for virtual screening in the chemical universe database GDB-13. *J. Am. Chem. Soc.* **131**, 8732–8733 (2009).
- Nakata, M. & Shimazaki, T. PubChemQC project: a large-scale first-principles electronic structure database for data-driven chemistry. *J. Chem. Inform. Model.* **57**, 1300–1308 (2017).
- Berman, H., Henrick, K. & Nakamura, H. Announcing the worldwide Protein Data Bank. *Nat. Struct. Mol. Bio.* **10**, 980 (2003).
- Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A. Machine learning for molecular and materials science. *Nature* **559**, 547–555 (2018).
- Groom, C. R. & Allen, F. H. The Cambridge structural database in retrospect and prospect. *Angew. Chem. Int. Ed.* **53**, 662–671 (2014).
- Groom, C. R., Bruno, I. J., Lightfoot, M. P. & Ward, S. C. The Cambridge structural database. *Acta Cryst. Sect. B* **72**, 171–179 (2016).
- Grazulis, S. et al. Crystallography Open Database—an open-access collection of crystal structures. *J. Appl. Cryst.* **42**, 726–729 (2009).
- Liu, F., Duan, C. & Kulik, H. J. Rapid detection of strong correlation with machine learning for transition-metal complex high-throughput screening. *J. Phys. Chem. Lett.* **11**, 8067–8076 (2020).
- 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.* **6**, 513–524 (2020).
- Nandy, A. et al. Machine learning accelerates the discovery of design rules and exceptions in stable metal–oxo intermediate formation. *ACS Catal.* **9**, 8243–8255 (2019).

- Vela, S., Fabrizio, A., Briling, K. R. & Corminboeuf, C. Learning the exciton properties of azo-dyes. *J. Phys. Chem. Lett.* **12**, 5957–5962 (2021).
- Blaskovits, J. T., Fumanal, M., Vela, S., Fabregat, R. & Corminboeuf, C. Identifying the trade-off between intramolecular singlet fission requirements in donor–acceptor copolymers. *Chem. Mater.* **33**, 2567–2575 (2021).
- Blaskovits, J. T., Fumanal, M., Vela, S. & Corminboeuf, C. Designing singlet fission candidates from donor–acceptor copolymers. *Chem. Mater.* **32**, 6515–6524 (2020).
- Vela, S., Krüger, C. & Corminboeuf, C. Exploring chemical space in the search for improved azoheteroarene-based photoswitches. *Phys. Chem. Chem. Phys.* **21**, 20782–20790 (2019).
- Brown, I. D. Recent developments in the methods and applications of the bond valence model. *Chem. Rev.* **109**, 6858–6919 (2009).
- Reeves, M. G., Wood, P. A. & Parsons, S. Automated oxidation-state assignment for metal sites in coordination complexes in the Cambridge Structural Database. *Acta Cryst. Sect. B* **75**, 1096–1105 (2019).
- Jablonka, K. M., Ongari, D., Moosavi, S. M. & Smit, B. Using collective knowledge to assign oxidation states of metal cations in metal–organic frameworks. *Nat. Chem.* **13**, 771–777 (2021).
- Shevchenko, A. P., Smolkov, M. I., Wang, J. & Blatov, V. A. Mining knowledge from crystal structures: oxidation states of oxygen-coordinated metal atoms in ionic and coordination compounds. *J. Chem. Inform. Model.* **62**, 2332–2340 (2022).
- Bruno, I. J., Shields, G. P. & Taylor, R. Deducing chemical structure from crystallographically determined atomic coordinates. *Acta Cryst. Sect. B* **67**, 333–349 (2011).
- Quirós, M., Gražulis, S., Girdzijauskaitė, S., Merkys, A. & Vaitkus, A. Using SMILES strings for the description of chemical connectivity in the Crystallography Open Database. *J. Cheminform.* **10**, 23 (2018).
- Shevchenko, A. P., Shabalin, A. A., Karpukhin, I., Yu & Blatov, V. A. Topological representations of crystal structures: generation, analysis and implementation in the TopCryst system. *Sci. Technol. Adv. Mater.* **2**, 250–265 (2022).
- Blatov, V. A., Shevchenko, A. P. & Proserpio, D. M. Applied topological analysis of crystal structures with the program package ToposPro. *Cryst. Growth Des.* **14**, 3576 (2014).
- Willett, P., Barnard, J. M. & Downs, G. M. Chemical similarity searching. *J. Chem. Inform. Comp. Sci.* **38**, 983–996 (1998).
- Balcells, D. & Skjelstad, B. B. tmQM dataset—quantum geometries and properties of 86k transition metal complexes. *J. Chem. Inform. Model.* **60**, 6135–6146 (2020).
- Sit, P. H. L., Car, R., Cohen, M. H. & Selloni, A. Simple, unambiguous theoretical approach to oxidation state determination via first-principles calculations. *Inorg. Chem.* **50**, 10259–10267 (2011).
- Gimferer, M., Aldossary, A., Salvador, P. & Head-Gordon, M. Oxidation state localized orbitals: a method for assigning oxidation states using optimally fragment-localized orbitals and a fragment orbital localization index. *J. Chem. Theory Comput.* **18**, 309–322 (2022).
- Thom, A. J. W., Sundstrom, E. J. & Head-Gordon, M. LOBA: a localized orbital bonding analysis to calculate oxidation states, with application to a model water oxidation catalyst. *Phys. Chem. Chem. Phys.* **11**, 11297–11304 (2009).
- Postils, V., Delgado-Alonso, C., Luis, J. M. & Salvador, P. An objective alternative to IUPAC's approach to assign oxidation states. *Angew. Chem. Int. Ed.* **57**, 10525–10529 (2018).
- Guha, R. et al. The Blue Obelisk—interoperability in chemical informatics. *J. Chem. Inform. Model.* **46**, 991–998 (2006).
- Björkman, T. CIF2Cell: generating geometries for electronic structure programs. *Comp. Phys. Commun.* **182**, 1183–1186 (2011).
- Hoppe, R. Effective coordination numbers (ECoN) and mean fictive ionic radii (MEFIR). *Z. Kristallogr. – Cryst. Mater.* **150**, 23–52 (1979).
- O'Keefe, M. & Brese, N. E. Atom sizes and bond lengths in molecules and crystals. *J. Am. Chem. Soc.* **113**, 3226–3229 (1991).
- Zimmermann, N. E. R. & Jain, A. Local structure order parameters and site fingerprints for quantification of coordination environment and crystal structure similarity. *RSC Adv.* **10**, 6063–6081 (2020).
- Kim, Y. & Kim, W. Y. Universal structure conversion method for organic molecules: from atomic connectivity to three-dimensional geometry. *Bull. Korean Chem. Soc.* **36**, 1769–1777 (2015).
- Godoy, F. et al. Synthesis, reactivity and molecular structure of phosphino tetramethyl cyclopentadienyl complex (η^5 - η^1 -C5Me4CH2PPh2)Re(CO)2. *Dalton Trans.* **16**, 3044–3051 (2009).
- Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inform. Comp. Sci.* **28**, 31–36 (1988).
- von Lilienfeld, O. A., Müller, K.-R. & Tkatchenko, A. Exploring chemical compound space with quantum-based machine learning. *Nat. Rev. Chem.* **4**, 347–358 (2020).

46. Ioannidis, E. I., Gani, T. Z. & Kulik, H. J. molSimplify: a toolkit for automating discovery in inorganic chemistry. *J. Comput. Chem.* **37**, 2106–2117 (2016).
47. Colón, Y. J., Gómez-Gualdrón, D. A. & Snurr, R. Q. Topologically guided, automated construction of metal–organic frameworks and their evaluation for energy-related applications. *Cryst. Growth Des.* **17**, 5801–5810 (2017).
48. Graser, J., Kauwe, S. K. & Sparks, T. D. Machine learning and energy minimization approaches for crystal structure predictions: a review and new horizons. *Chem. Mater.* **30**, 3601–3612 (2018).
49. Beran, G. J. O. Modeling polymorphic molecular crystals with electronic structure theory. *Chem. Rev.* **116**, 5567–5613 (2016).
50. Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
51. Huang, B. & von Lilienfeld, O. A. Quantum machine learning using atom-in-molecule-based fragments selected on the fly. *Nat. Chem.* **12**, 945–951 (2020).
52. Christensen, A. S. et al. QML: A Python Toolkit for Quantum Machine Learning. (2017).

ACKNOWLEDGEMENTS

This research was supported by the National Centre of Competence in Research (NCCR) MARVEL, a NCCR funded by the Swiss National Science Foundation (grant number 182892). The NCCR Catalysis (grant number 180544) of SNSF is also acknowledged for financial support of R.L.

AUTHOR CONTRIBUTIONS

C.C. and S.V. conceived the project. S.V., R.L., and Y.C. wrote the code *cell2mol*, constructed the databases, designed the ML models, implemented the BVS method, searched the literature, and performed all analyses. All authors discussed the results and assisted with the manuscript preparation.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41524-022-00874-9>.

Correspondence and requests for materials should be addressed to Clémence Corminboeuf.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022