Received: 2 October 2019 | Revised: 19 May 2020 | Accepted: 11 June 2020

DOI: 10.1002/qua.26382

FULL PAPER

International Journal of Quantum Chemistry WILEY

Turning chemistry into information for heterogeneous catalysis

Sergio Pablo-García[1] | Moises Álvarez-Moreno[1,2] | Núria López[1]

[1]The Barcelona Institute of Science and Technology, Institute of Chemical Research of Catalonia, ICIQ, Tarragona, Spain

[2]Department of Physical and Inorganic Chemistry, Universitat Rovira i Virgili, Tarragona, Spain

Correspondence
Sergio Pablo-García or Núria López,
The Barcelona Institute of Science and Technology, Institute of Chemical Research of Catalonia, ICIQ, Avinguda dels Països Catalans 16, 43007 Tarragona, Catalonia, Spain.
Email: spgarcia@iciq.es, nlopez@iciq.es

Funding information
Ministerio de Ciencia e Innovación, Grant/Award Number: RTI2018-101394-B-I00

Abstract

The growing generation of data and their wide availability has led to the development of tools to produce, analyze, and store this information. Computational chemistry studies, especially catalytic applications, often yield a vast amount of chemical information that can be analyzed and stored using these tools. In this manuscript, we present a framework that automatically performs a fully automated procedure consisting of the transfer of an adsorbate from a known metal slab to a new metal slab with similar packing. Our method generates the new geometry and also performs the required calculations and analysis to finally upload the processed data to an online database (ioChem-BD). Two different implementations have been built, one to relocate minimum energy point structures and the second to transfer transition states. Our framework shows good performance for the minimum point location and a decent performance for the transition state identification. Most of the failures occurred during the transition state searches and needed additional steps to fully complete the process. Further improvements of our framework are required to increase the performance of both implementations. These results point to the *avoidhuman* path as a feasible solution for studies on very large systems that require a significant amount of human resources and, in consequence, are prone to human errors.

KEYWORDS

avoidhumans, database, FAIR data, heterogeneous catalysis, information science

1 | INTRODUCTION

Nowadays, computational chemistry is ubiquitous and has applications in chemistry, biology, physics, materials science, and nanotechnology. As the access to massive computers and robust codes[1] extends worldwide, databases for molecules, nanostructures and materials containing structural data, spectroscopic fingerprints,[2] and general properties can be easily generated. The ultimate applications of these databases can vary from environmental detection through spectroscopy to data mining for materials in catalysis and electrocatalysis.[3] Yet, most of the purpose-oriented calculations are not saved (this is the general case in materials and heterogeneous catalysis) or are only presented as lengthy xyz coordinate listings in error-prone supplementary files. Only lately, the relevance of keeping these data in the form of databases has been acknowledged.[4] Most of the systems, however, have emerged in materials science in projects such as the Materials Project,[5,6] NoMaD,[7] Materials Cloud,[8] and Computational Materials Repository.[9] Data are mostly unlinked to the corresponding works, and thus, the traceability (who, when, what) and fairness (findable, accessible, interoperable, reusable) are lost.[10]

Human factors pose an additional problem. Researchers have to perform very routine tasks; therefore, they can make mistakes, and the generated dense datasets often have multiple deficiencies. Human factors include, for example, cognitive functions (such as attention, detection,

perception, judgment, and reasoning, including heuristics and biases) and decision making. Each of these is further divided into subcategories. These issues are particularly problematic when statistical learning techniques[11–21] are applied; sparse datasets are biased toward a particular type of successful event, exactly what statistical learning algorithms need to avoid to ensure their robustness. Repetitiveness has been addressed by different groups by using scripts with different levels of sophistication. However, the emergence of new frameworks can ease the tedious tasks and generate/check/upload to a database significant data blocks of the phase spaces and provide the right tools into the automation concepts.[22–25] This has lead the (pun) idea of *avoidhumans* in the data production. Generation is still one of the key steps. Some very recent efforts have been made to automate the identification of adsorption sites in metal surfaces and to create new structures with different combinations of sites and adsorbates.[26–28] Statistical learning techniques have also been applied to obtain an estimate for the adsorption energies.[29,30] Our work tries to use structure inheritance between different metals not only to automate the study of reaction networks in different metals and alike materials (like oxides) but also to reduce the explicit density functional theory (DFT) calculation time for adsorbates (a relatively straightforward task) and, most relevantly, for transition states (TSs) for a real catalytic problem. Diagnose exceptions and analysis will be the focus of research in computational chemistry in the coming years. This will increase our abilities to find outliers that can be crucial to performance (and identification of new families of molecules and compounds with particularly appealing properties); to refine the analytics; to incorporate graph theory[31] or other encodings, such as SMILES[32]; and to be able to transfer active patterns irrespective of the nature of the compound (solid, enzymatic, molecular).

The ways of acquiring information have changed along with data generation and analysis. New editorial platforms such as Authorea can also integrate the advances of these systematic approaches. The process of reading documents has changed drastically since the establishment of the World Wide Web and the possibility of several instances running simultaneously. Reading in the 21st century is a completely different experience than it has been for at least 500 years as the meta- and linked data are accessible and are consulted almost simultaneously with the primary source. New ways of reading can thus now benefit from interactive viewers that can integrate the content and improve the visualization of complex information (for instance, 3D).

Our manuscript tries to address all these new challenges in computational chemistry with the overall idea of transforming the results into a true seamless information science that can be interactively read, dynamically searched, analyzed, and mined while ensuring the transferability among different fields through metadata storage. Hence, the combination of Fireworks,[33] ioChem-BD,[34,35] and Authorea[36] constitutes a privileged platform.

2 | COMPUTATIONAL DETAILS

DFT simulations have been performed using Vienna ab-initio Simulation Package (VASP).[37,38] Generalized Gradient Approximation with the Perdew-Burke-Ernzerhof functional has been used to obtain the exchange-correlation energy term.[39] Valence electrons have been represented using the projector augmented wave[40] with a kinetic energy cutoff of 450 eV. A Γ-centered mesh has been built using the Monkorst-Pack method[41] to generate the k-points. The improved dimer method has been used to locate the TSs.[42]

Fireworks[33] is an open-source software package that allows managing, building, and running workflows. In this project, it has been used to build and configure the ties between the different steps of our calculations. Fireworks also allows one to track the status of active workflows and stops the process if one of the steps fails, allowing one to restart the process completely or to continue the step after the application of the needed changes. Software written in Python and Bash has been developed in our lab and used to script the preparation, transfer, and checking processes. Developed Python libraries and scripts focus on geometry manipulation and input/output parsing, while Bash scripts manage the files related to the calculations and control the execution of VASP.

The framework has been tested using a reaction network composed of nine different metals, seven of them with p(3x3)-(111) fcc packing and two with p(3x3)-(0001) hcp packing. One of the hcp metals serves as the ansatz host for the rest of the metals. Both packings share similar adsorption sites in their surfaces. The two lowest layers have been frozen, and a vacuum of 15 Å has been applied to all the metal slabs.

A total of 12 organic intermediates with the chemical formula $C_1Br_xH_y$ and eight TSs have been evaluated with our method. All the intermediates belong to the same reaction network; the TSs are all the possible elemental steps that link all our intermediates. For the minimum energy relaxations (MERs), different adsorption sites have been calculated for every metal/species combination, while only one adsorption site has been tested for the TSs. The species were first obtained manually for the hcp metal and then recalculated for the other metals by using our framework.

The figures included in this paper were made with Cytoscape, 3Dmol and Plotly.[43–45]

3 | RESULTS AND DISCUSSION

Full mechanistic DFT studies on the decomposition of organic molecules are among the most challenging processes in heterogeneous catalysis due to the number of elemental steps and intermediate species comprising the network, as well as to the subsequent screening that the

calculation of these reactions over a set of metal slabs require.[46] The size of these networks are tightly bonded with the number of carbons of the organic species, and for a significant number of carbons, the systems require a massive amount of time and resources to be evaluated. To address this issue, some databases have been designed to store and ease the access to these steps with the aim of recycling old calculations and reducing the required steps when studying these networks.[47]

We propose an automated procedure to overcome the drawbacks associated with the repetitive processes required for the study of large reaction networks. We have built a framework that combines Fireworks, VASP, ioChem-BD, and ad-hoc developed software to fully automate the study of reaction networks over different metals. This framework allows transferring all the relaxed species and TSs obtained for a metal slab to the slab of a different metal, as well as preparing and performing the DFT calculations by using the generated geometries for the new slab automatically. The simplicity of both the C_1 species and the pure metal slabs, together with the complexity degree added by the inclusion of a halogen, give us an optimal scenario to benchmark the framework.

3.1 | Transfer algorithm

The process of generating guess geometries for intermediates is one of the most tedious and time-consuming tasks, particularly when dealing with large reaction networks, for example, C_6 sugar alcohol decomposition consists of 10^6 intermediates.[48] Inheriting previously obtained structures from similar calculations eases the problem, but it still results in a repetitive routine problem that can be fully automated.

A transfer algorithm that performs the relocation of an adsorbed molecule in a metal slab to a similar metal slab has been developed. Automatic transfer of adsorbed species between similar metallic surfaces allows not only saving a considerable amount of time during the geometry production but also classifying the generated geometries accurately.

When there are two different metal slabs, a host with an adsorbed molecule (1a) and a guest consisting of an empty metal slab (1b), the transfer algorithm works as follows: First, it identifies the layers (2) for both slabs and selects the highest layer (3). Then, it searches for all the possible adsorption sites (4) on both surfaces. The site is found by triangulation of the different metal centers. Once the adsorption sites have been identified, the algorithm associates the nearest adsorption site with the lowest atom of the adsorbed molecule in the host slab. (5) In the next step, the adsorption length is computed by using the distance between the assigned site and the lowest atom of the adsorbate. However, for some molecules, the lowest atom (*z*-axis position) is not perfectly aligned with the adsorption site. To overcome this issue, the deviation of the *z* vector between both is also computed in this step (6). Finally, the algorithm transfers the molecule to a similar site in the guest surface, taking into account the adsorption site type and maintaining the adsorption length (7). As an optional step, the algorithm can be set to identify the nearest adsorption site of the second lowest (5o) atom and compute the angle between both (6o); this information is then used to rotate the molecule around the *z*-axis of the lowest atom to preserve the original alignment of the molecule (8). Figure 1 depicts the procedure of the algorithm.

The transfer algorithm applies different methods to find the possible adsorption sites. For the fcc and hcp holes detection, the Voronoi tetrahedron method[49] is used to compute the bonds between the atoms of the upper layer and then to search for cycles of three atoms. Differentiation between hcp and fcc holes is achieved by projecting the triangle formed by the cycles in the lower layer and searching for atoms inside this space. Once the bonds are defined, the bridge and top positions are easy to find.

3.2 | Workflow design

The geometries generated using the transfer algorithm are not optimized and require further DFT calculations to obtain a full description of the reaction network. To address this problem, two different workflows have been developed and embedded in the framework: one for the MER and another for the TS search. Both workflows use the transfer algorithm as the first step to generate a new metal slab with the desired adsorbate; then, different Bash scripts are prepared and run the pertinent DFT calculations using VASP.

While the workflow to search for relaxed geometries only generates a single geometry, the TS workflow generates three unique geometries with different rotations along the *z*-axis of the lowest atom. A partial minimum energy search for a small number of ionic steps is performed for the three structures, and the atoms of the molecule with high initial forces are allowed to relax, thus preventing the TS to fall to a minimum energy point. The best candidate among the relaxed structures is selected using a lowest energy criterion; in the next step, the chosen candidate is used as the starting geometry in an improved dimer method calculation. Selection steps are unnecessary for the MER workflow due to the efficiency of the algorithms that relax the geometry to the minimum point energy compared with the transition search ones. Thus, the generated geometry for MER workflow is directly used as starting point for the minimum energy search calculations.

Different errors can occur during the calculation steps; besides, the geometry obtained after the calculation may be chemically meaningless, and therefore, it is important to manage errors with care. To address these issues, a checking algorithm is applied to the results of the calculations in order to search for inconsistencies through the calculation steps. In addition, a bond identification algorithm is used to verify that no bond breaking occurs during the relaxations.

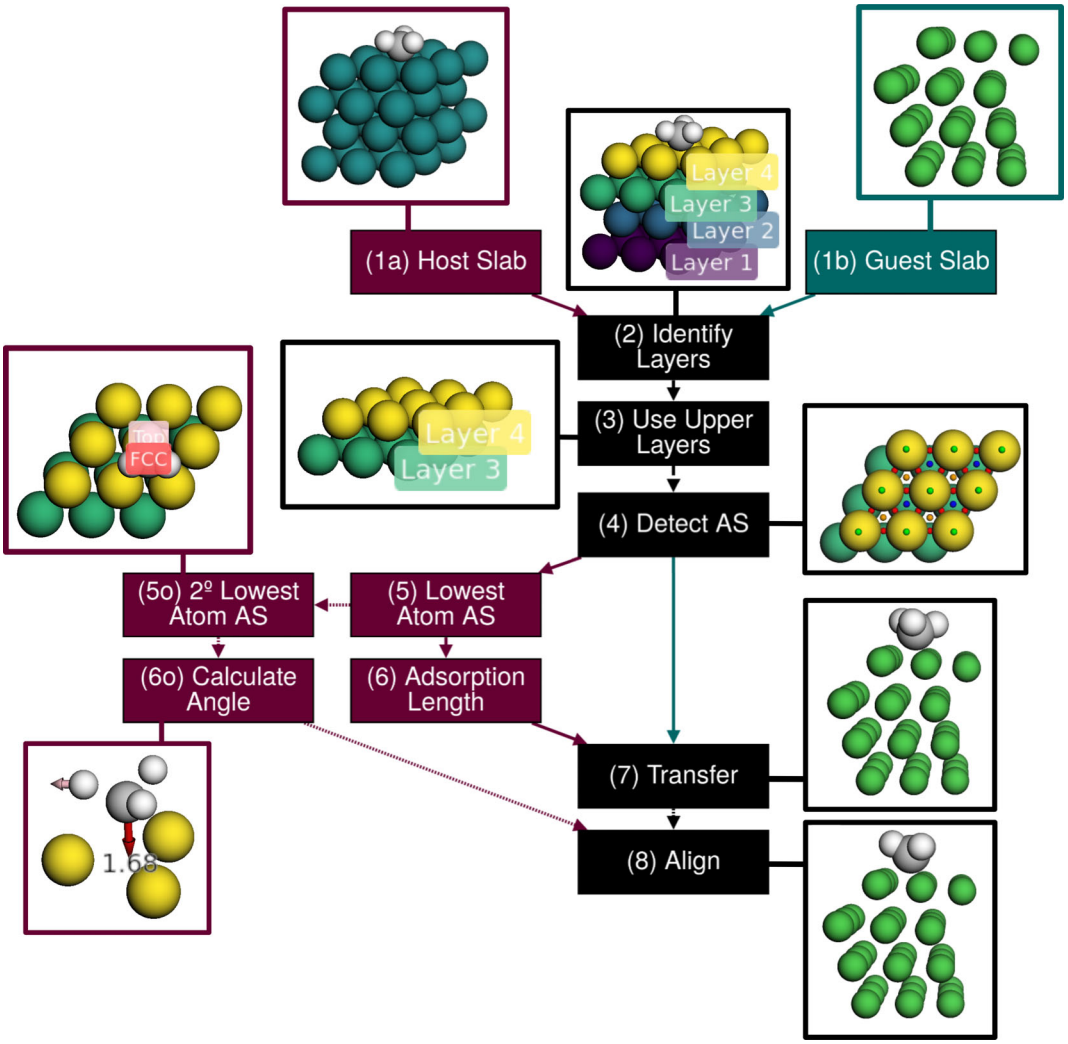

FIGURE 1 Interactive diagram of the steps that the transfer algorithm performs. While some of the steps only use the data from the host slab, the first (1) to (5) steps are applied concurrently to both slabs

This algorithm, which detects spurious, nonvalid broken adsorbates, was implemented at the checking stage. The process splits the structure between the (metal) surface and the adsorbed molecules. Then, the bonds between the atoms of the adsorbates are detected using the Voronoi tetrahedra algorithm[49] (with a cutoff radius) and converted to a graph, and then, the number of disconnected graphs is computed. The difference between the initial and final number of disconnected graphs illustrates whether the adsorbate has broken during optimization. If the result passes the check, a frequency calculation is launched.

After a few tests, improvement has been integrated into the transfer algorithm. The bond recognition was reused to analyze the bonds of the geometries obtained from the MER workflows. As a result, a list of the average distance between the metal surface and the lower atom was obtained for each metal surface. The difference between these distances was used to correct the adsorption distance of some of the failed MER workflows and all of the TS workflows.

3.3 | Storing data

Due to the large number of individual results that compose complex reaction networks, it is mandatory to compile, sort, and store the results. refs. [34, 35] provides the essential tools to perform the last step of our project: to convert our results into organized data. The shell client of ioChem-BD allows an easily upload of the generated output files to a private server; once these are uploaded, it transforms, parses, and stores the results to ensure a clear representation and easy online access to the data. It can also generate molecular labels as SMILES.

Consequently, as the final step of both previously described workflows, the results of the DFT calculations, as well as the frequency calculations performed at the end of the workflows, are uploaded to ioChem-BD. For the TS workflow, the data generated by the dimer method are uploaded, whereas for the MER workflow, the data from the relaxation method are uploaded.[50,51]

Once all the data are processed, they can be published in the public repository of the ioChem-BD server, and then, the obtained information can be shared with other researchers in accordance with the FAIR principles. Moreover, ioChem-BD generates interactive figures that improve the understanding of geometries and reactivity for complex adsorbed molecules when working in multidisciplinary environments (eg, with experimental groups).

Datasets can also be published with an embargo option; the dataset is published in a private repository, and both a DOI and a Reviewer Link are generated. This link allows the coworkers, editor, and reviewers to inspect the dataset before making it public. Once the associated manuscript is published, the dataset is synchronized with the public repository, making it accessible to everyone through the dataset's DOI. The data can then be accessed via DOI or by searching directly in the platform by author, date, SMILES, or chemical formula.

When a calculation is uploaded to ioChem-BD, additional metadata are generated and stored within the calculation; a trustworthy fingerprint of every calculation is also created and deposited in the system. Table 1 includes the minimum parameters that are saved when a calculation is uploaded to ioChem-BD. Two different schemes are used to generate the content tree: the Dublin Core standard[52] and the Chemical Custom format created for ioChem-BD to append additional data.

While the output file from VASP is processed and converted to our custom .cml format, the raw input files are stored as they are. To optimize space usage, the particular electronic and ionic iterations are not parsed, and only the final structure is stored. Inclusion of the raw input files and the VASP version used allows reproducing every calculation of the dataset accurately.

Additional information on the structure of the different calculations or just for clarity can also be added. ioChem-BD offers several options to include this information, and notes can be added to its description. However, this information is rather complex in some cases and requires a different format. When this happens, ioChem-BD offers the possibility of attaching raw files together with the input files. Both the description and the attachments are available once the calculation is published in the ioChem-BD server.

3.4 | Test and results

Both workflows yield reasonable results and apply the procedure for the different species (see Figure 3). Almost 88% of the MERs and 56% of the TSs converge at the first attempt. Additional steps are required for 0.1% of the relaxed models and 27% of the TSs. The rest require manual preparation and supervision for them to terminate successfully. The exhaustion of the number of cycles was the main cause of the failures for MER, and the solution was to add about 10% ionic cycles (not completely used in all cases). This can thus be directly introduced in alternative network studies for metals and alloys. For TSs, there were two issues: (a) the same as for MER, insufficient ionic steps in the algorithm, and this was sorted in the same manner and was by far the most common issue; (b) alternative convergence was not achieved, and thus, the initial seed was not taken from the reference host (Ru) but from a metal with chemical properties closer to the running TS search (Pt was inherited from Ir).

In the case of relaxations, most of the unobtainable structures fall into different adsorption sites during the calculation, while only a few of them end the calculation with a broken bond. On the other hand, all the unobtainable TSs end with a broken bond, and more precise methods such as the Nudged-Elastic Band[53,54] are required to obtain the geometries.

This automatization procedure has been used to generate the databases for further machine learning studies.[55]

TABLE 1 Minimum metadata stored in ioChem-BD for each calculation

Dublin core (dc)	Chemical custom
contributor.author	program.name
contributor.other	program.version
date.accessioned	program.other
date.available	method
date.created	shelltype
date.issued	energy.value
identifier.uri	energy.units
description	formula.generic
publisher	hassolvent
relation.ispartof	hasvibrationalfrequencies
rights	numberofjobs
rights.uri	hasmolecularorbital
title	
type	
date.updated	

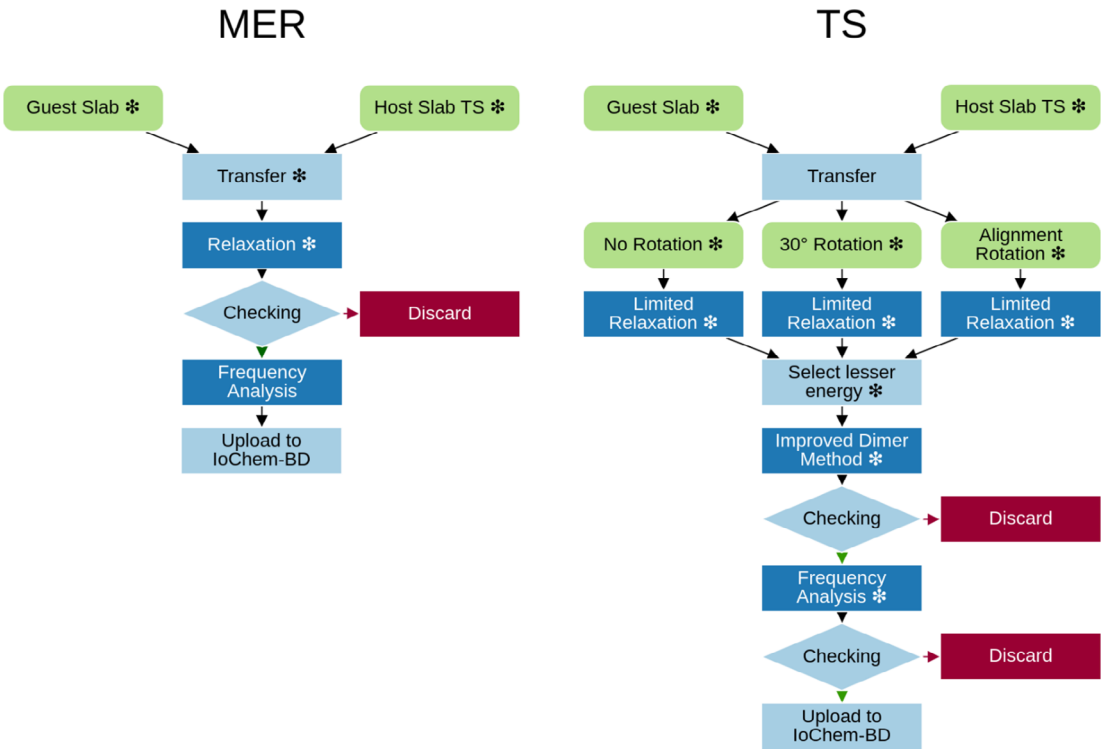

FIGURE 2 Interactive flux diagram showing the required steps for the relax and TS processes. DFT calculations are colored in dark blue and scripts in light blue. Clicking the different points marked with (*) will show the status of the structure at this point (interactive figure only)

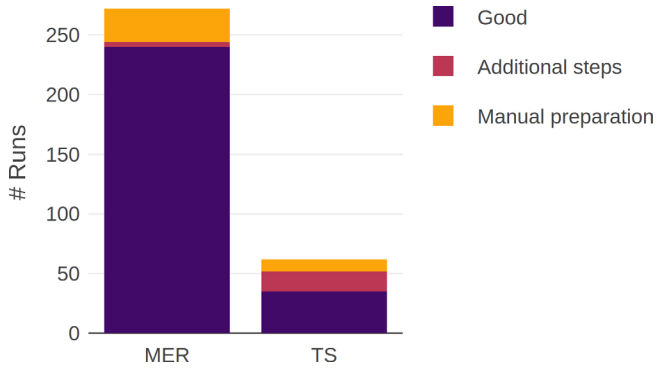

FIGURE 3 Efficiency of the MER and TS workflows. "Good" points to regularly terminated calculations, "Additional steps" imply that error checking routines have identified a deficiency in the calculation (ie, maximum convergence steps reached), and "Manual preparation" points out that the calculation failed and requires human intervention (ie, bond breaking)

3.5 | Integration with authorea

The Authorea[36] platform allows easy integration of the data stored in ioChem-BD. For instance, the reaction networks and particular structures can be directly linked in the database. In our case, the structure for the CH_3 on two different surfaces, Ni and Ru, can be retrieved from Guest Slab and Host Slab and the TS for decomposition from Transition State.

4 | CONCLUSIONS

We have proved that our framework automates two different kinds of molecular transfers through similar metals. Although our method is not yet able to fully automate the entire process successfully, it is possible to classify the different error cases obtained during our study and incorporate the solutions as additional steps for our workflows. In addition, coupling our framework to the Authorea and ioChem-BD tools has the following advantages: (a) it enables establishing a seamless link between the computed data and the manuscript and links the corresponding structures interactively, thus avoiding tedious and error-prone supporting information; (b) this is particularly attractive for complex databases with massive reaction networks and/or several material compositions; (c) the workflow reduces the computing time, systematizes the nomenclature and

labeling of the different species, reduces the chaos, and increases transferability; (d) the metadata directly embedded in ioChem-BD is made transparent through Authorea, which can improve the design (and self-definition) of the working flows that could potentially include data analysis through coupled machine learning algorithms; and (e) the data curation is thus directly enforced by this procedure.

ACKNOWLEDGMENTS

We are thankful to the "Ministerio de Ciencia e Innovación" RTI2018-101394-B-I00 and the Barcelona Supercomputing Center (BSC-RES) for providing generous computer resources. We thank Ms Vendrell for carefully reading the manuscript.

AUTHOR CONTRIBUTIONS

Sergio Pablo-García: Data curation; investigation; methodology; validation; visualization; writing-original draft; writing-review and editing. **Moises Álvarez-Moreno:** Conceptualization; data curation; investigation; methodology; visualization; writing-original draft; writing-review and editing. **Núria López:** Conceptualization; funding acquisition; investigation; supervision; writing-original draft; writing-review and editing.

DATA ACCESSIBILITY

The data that support the findings of this study are openly available in ioChem-DB at doi: 10.19061/iochem-bd-1-137 (Figure 2) and doi: 10.19061/iochem-bd-1-150 (Figure 3).

PEER REVIEW DETAILS

Open Peer Reviewer Details for this article are openly available here:

- Peer Reviewer Report #1 DOI: 10.22541/au.159182659.98933679.
- Peer Reviewer Report #2 DOI: 10.22541/au.159182658.82141376.
- Editor's Comment DOI: 10.22541/au.159182678.86670339.
- Author Response DOI: 10.22541/au.159188746.63115568.

ORCID

Sergio Pablo-García https://orcid.org/0000-0002-3327-9285

REFERENCES

[1] K. Lejaeghere, G. Bihlmayer, T. Björkman, P. Blaha, S. Blügel, V. Blum, D. Caliste, I. E. Castelli, S. J. Clark, A. Dal Corso, et al., *Science* **2016**, *351*, aad3000.
[2] S. Grimme, C. Bannwarth, S. Dohm, A. Hansen, J. Pisarek, P. Pracht, J. Seibert, F. Neese, *Angew. Chem. Int. Ed.* **2017**, *56*, 14763.
[3] J. R. Kitchin, *Nat. Catal.* **2018**, *1*, 230.
[4] C. Bo, F. Maseras, N. López, *Nat. Catal.* **2018**, *1*, 809.
[5] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, K. A. Persson, *APL Mater.* **2013**, *1*, 011002.
[6] Materials Project, https://materialsproject.org/ (accessed: January 2020).
[7] NoMaD Repository, https://nomad-coe.eu/ (accessed: January 2020).
[8] Materials Cloud, https://www.materialscloud.org/home (accessed: January 2020).
[9] Computational Materials Repository, https://cmr.fysik.dtu.dk/ (accessed: January 2020).
[10] M. D. Wilkinson, M. Dumontier, I. J. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. G. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. C. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S. A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, B. Mons, *Sci. Data* **2016**, *3*, 160018.
[11] R. García-Muelas, N. López, *Nat. Commun.* **2019**, *10*. http://dx.doi.org/10.1038/s41467-019-12709-1.
[12] A. Bruix, J. T. Margraf, M. Andersen, K. Reuter, *Nat. Catal.* **2019**, *2*, 659.
[13] L. Turcani, R. L. Greenaway, K. E. Jelfs, *Chem. Mater.* **2018**, *31*, 714.
[14] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, A. Walsh, *Nature* **2018**, *559*, 547.
[15] G. Gryn'ova, K.-H. Lin, C. Corminboeuf, *J. Am. Chem. Soc.* **2018**, *140*, 16370.
[16] Z. W. Ulissi, A. J. Medford, T. Bligaard, J. K. Nørskov, *Nat. Commun.* **2017**, *8*. http://dx.doi.org/10.1038/ncomms14621.
[17] R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams, A. Aspuru-Guzik, *ACS Cent. Sci.* **2018**, *4*, 268.
[18] P. SchlexerLamoureux, K. T. Winther, J. A. GarridoTorres, V. Streibel, M. Zhao, M. Bajdich, F. Abild-Pedersen, T. Bligaard, *ChemCatChem* **2019**, *11*, 3581.
[19] B. Meyer, B. Sawatlon, S. Heinen, O. A. von Lilienfeld, C. Corminboeuf, *Chem. Sci.* **2018**, *9*, 7069.
[20] A. Nandy, J. Zhu, J. P. Janet, C. Duan, R. B. Getman, H. J. Kulik, *ACS Catal.* **2019**, *9*, 8243.

[21] P. Z. Moghadam, S. M.J. Rogge, A. Li, C.-M. Chow, J. Wieme, N. Moharrami, M. Aragones-Anglada, G. Conduit, D. A. Gomez-Gualdron, V. Van Speybroeck, D. Fairen-Jimenez, *Matter* **2019**, *1*, 219. http://dx.doi.org/10.1016/j.matt.2019.03.002.
[22] A. H. Larsen, J. J. Mortensen, J. Blomqvist, I. E. Castelli, R. Christensen, M. Dułak, J. Friis, M. N. Groves, B. Hammer, C. Hargus, et al., *J. Phys. Condens. Matter* **2017**, *29*, 273002.
[23] G. Pizzi, A. Cepellotti, R. Sabatini, N. Marzari, B. Kozinsky, *Comput. Mater. Sci.* **2016**, *111*, 218.
[24] Open Babel: The Open Source Chemistry Toolbox, http://openbabel.org/wiki/Main_Page (accessed: January 2020).
[25] P. S. Gromski, J. M. Granda, L. Cronin, *Trends Chem.* **2020**, *2*, 4. http://dx.doi.org/10.1016/j.trechm.2019.07.004.
[26] J. R. Boes, O. Mamun, K. Winther, T. Bligaard, *Chem. A Eur. J.* **2019**, *123*, 2281.
[27] K. Tran, A. Palizhati, S. Back, Z. W. Ulissi, *J. Chem. Inf. Model.* **2018**, *58*, 2392.
[28] J. H. Montoya, K. A. Persson, *npj Computational Materials* **2017**, *3*. http://dx.doi.org/10.1038/s41524-017-0017-z.
[29] Z. W. Ulissi, M. T. Tang, J. Xiao, X. Liu, D. A. Torelli, M. Karamad, K. Cummins, C. Hahn, N. S. Lewis, T. F. Jaramillo, K. Chan, J. K. Nørskov, *ACS Catal.* **2017**, *7*, 6600.
[30] D. P. Tabor, L. M. Roch, S. K. Saikin, C. Kreisbeck, D. Sheberla, J. H. Montoya, S. Dwaraknath, M. Aykol, C. Ortiz, H. Tribukait, C. Amador-Bedolla, C. J. Brabec, B. Maruyama, K. A. Persson, A. Aspuru-Guzik, *Nat. Rev. Mater.* **2018**, *3*, 5.
[31] P. Bjørn Jørgensen, E. Garijo del Río, M. N. Schmidt, K. Wedel Jacobsen, *arXiv e-prints* **2019**, *arXiv:1905.06048*.
[32] D. Weininger, *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31.
[33] A. Jain, S. P. Ong, W. Chen, B. Medasani, X. Qu, M. Kocher, M. Brafman, G. Petretto, G.-M. Rignanese, G. Hautier, D. Gunter, K. A. Persson, *Concurr. Comput. Pract. Exp.* **2015**, *27*, 5037.
[34] M. Álvarez-Moreno, C. de Graaf, N. Lopez, F. Maseras, J. M. Poblet, C. Bo, *J. Chem. Inf. Model.* **2014**, *55*, 95.
[35] ioChem-BD, https://www.iochem-bd.org/ (accessed: January 2020).
[36] Open Research Collaboration and Publishing – Authorea, https://authorea.com/ (accessed: January 2020).
[37] G. Kresse, J. Furthmüller, *Comput. Mater. Sci.* **1996**, *6*, 15.
[38] G. Kresse, J. Furthmüller, *Phys. Rev. B* **1996**, *54*, 11169.
[39] J. P. Perdew, K. Burke, M. Ernzerhof, *Phys. Rev. Lett.* **1996**, *77*, 3865.
[40] P. E. Blöchl, *Phys. Rev. B* **1994**, *50*, 17953.
[41] H. J. Monkhorst, J. D. Pack, *Phys. Rev. B* **1976**, *13*, 5188.
[42] A. Heyden, A. T. Bell, F. J. Keil, *J. Chem. Phys.* **2005**, *123*, 224101.
[43] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, T. Ideker, *Genome Res.* **2003**, *13*, 2498.
[44] N. Rego, D. Koes, *Bioinformatics* **2014**, *31*, 1322.
[45] Plotly Technologies Inc. Collaborative data science, https://plot.ly/ (accessed: January 2020).
[46] R. G. Muelas, Q. Li, N. López, **2017**, https://doi.org/10.19061/iochem-bd-1-37.
[47] CatApp database, https://cmr.fysik.dtu.dk/catapp/catapp.html (accessed: January 2020).
[48] J. E. Sutton, D. G. Vlachos, *Chem. Eng. Sci.* **2015**, *121*, 190.
[49] O. Isayev, C. Oses, C. Toher, E. Gossett, S. Curtarolo, A. Tropsha, *Nat. Commun.* **2017**, *8*, 15679.
[50] S. Pablo-García, Automation Transfer Dataset **2019**, https://doi.org/10.19061/iochem-bd-1-137
[51] S. Pablo-García, Full Dataset **2020**, https://doi.org/10.19061/iochem-bd-1-150
[52] DCMI, https://www.dublincore.org/ (accessed: January 2020).
[53] G. Henkelman, H. Jónsson, *J. Chem. Phys.* **2000**, *113*, 9978.
[54] G. Henkelman, B. P. Uberuaga, H. Jónsson, *J. Chem. Phys.* **2000**, *113*, 9901.
[55] A. J. Saadun, S. Pablo-Garcia, V. Paunovic, Q. Li, A. Sabadell-Rendón, K. Kleemann, F. Krumeich, N. Lopez, J. Pérez-Ramírez, *ACS Catal.* **2020**, *10*, 6129.

How to cite this article: Pablo-García S, Álvarez-Moreno M, López N. Turning chemistry into information for heterogeneous catalysis. *Int J Quantum Chem.* 2021;121:e26382. https://doi.org/10.1002/qua.26382