Turning chemistry into information for heterogeneous catalysis

Sergio Pablo-Garca[1,2], Moises lvarez-Moreno[1,3], and Nria Lpez[1,2]

[1]Institute of Chemical Research of Catalonia, ICIQ, Av. Pasos Catalans 16, 43007 Tarragona, Catalonia, Spain
[2]The Barcelona Institute of Science and Technology, BIST
[3]Department of Physical and Inorganic Chemistry, Universitat Rovira i Virgili, C/Marcell Domingo s/n, 43007 Tarragona, Catalonia, Spain

May 19, 2020

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 and 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 full automated procedure consisting in 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 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.

1 Introduction

Computational chemistry is nowadays ubiquitous and has applications in Chemistry, Biology, Physics, Materials Science and Nanotechnology. As the access to massive computers and robust codes (Lejaeghere et al., 2016) extends worldwide, databases for molecules, nanostructures and materials containing structural data, spectroscopic fingerprints (Grimme et al., 2017) 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 (Kitchin, 2018). And 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 this data in the form of databases has been acknowledged (Bo et al., 2018). Most of the systems, though, have emerged in Materials Science in projects such as the Materials Project (Jain et al., 2013; *Materials Project*, n.d.), NoMaD (*NoMaD Repository*, n.d.), Materials Cloud (*Materials Cloud*, n.d.) and Computational Materials Repository (*Computational Materials Repository*, n.d.). Data are mostly unlinked to the corresponding works and thus the traceability (who, when, what) and fairness (findable, accessible, interoperable, reusable) are lost (Wilkinson et al., 2016).

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, judgement and reasoning, including heuristics and biases), and decision making. Each of these is further divided into sub-categories. These issues are particularly problematic when statistical learning techniques (Garca-Muelas & Lpez, 2019; Bruix et al., 2019; Turcani et al., 2018; Butler et al., 2018; Gryn'ova et al., 2018; Ulissi et al., 2017; Gómez-Bombarelli et al., 2018; Schlexer Lamoureux et al., 2019; Meyer et al., 2018; Nandy et al., 2019; Moghadam et al., 2019) are applied; sparse datasets are biased towards 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 level of sophistication. However, the emergence of new frameworks that 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 (Larsen et al., 2017; Pizzi et al., 2016; *Open Babel: Te Open Source Chemistry Toolbox*, n.d.; Gromski et al., 2019). 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 (Boes et al., 2019; Tran et al., 2018; Montoya & Persson, 2017). Statistical learning techniques have also been applied to get an estimate for the adsorption energies (Ulissi et al., 2017; Tabor et al., 2018). 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 calculation time for adsorbates (a relatively straightforward task) and, most relevantly, for transition states for a real catalytic problem. Diagnose exceptions and analysis will be the focus of the 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 (Bjørn Jørgensen et al., 2019) or other encodings, like SMILES (Weininger, 1988), 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 like 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 and with the overall idea of transforming the results into a true seamless information science that can be interactively read, dynamically searched, analyzed and mined, whilst ensuring the transferability among different fields through metadata storage. Hence, the combination of Fireworks (Jain et al., 2015), ioChem-BD (Álvarez-Moreno et al., 2014; *ioChem-BD*, n.d.) and Authorea (*Open Research Collaboration and Publishing - Authorea*, n.d.) constitutes a privileged platform.

2 Computational Details

Density functional theory (DFT) simulations have been performed using Vienna ab-initio Simulation Package (VASP) (Kresse & Furthmüller, 1996; Kresse & Furthmüller, 1996). Generalized Gradient Aproximation with the Perdew-Burke-Ernzerhof functional (GGA-PBE) have been used to obtain the exchange-correlation energy term (Perdew et al., 1996). Valence electrons have been represented using the Projector Augmented Wave (PAW) (Blöchl, 1994) with a kinetic energy cutoff of 450 eV. A Γ-centered mesh has been built using the Monkorst-Pack method (Monkhorst & Pack, 1976) to generate the k-points. The Improved Dimer Method

has been employed to locate the transition states (Heyden et al., 2005).

Fireworks (Jain et al., 2015) 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 to track the status of active workflows and stops the process if one of the steps fails, allowing 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 9 different metals, 7 of them with p(3x3)-(111) fcc packing and 2 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 metal slabs.

A total of 12 organic intermediates with the chemical formula $C_1Br_xH_y$ and 8 transition states have been evaluated with our method. All the intermediates belong to the same reaction network; the transition states are all the possible elemental steps that link all our intermediates. For the minimum energy relaxations, different adsorption sites have been calculated for every metal/species combination, while only one adsorption site has been tested for the transition states. 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 (Shannon et al., 2003; Rego & Koes, 2014; Inc., 2015).

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 amount of elemental steps and intermediate species comprised in the network, as well as to the subsequent screening that the calculation of these reactions over a set of metal slabs require (Muelas et al., 2017). 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 (*CatApp database*, n.d.).

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 transition states 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 a optimal scenario to benchmark the framework.

3.1 Transfer algorithm

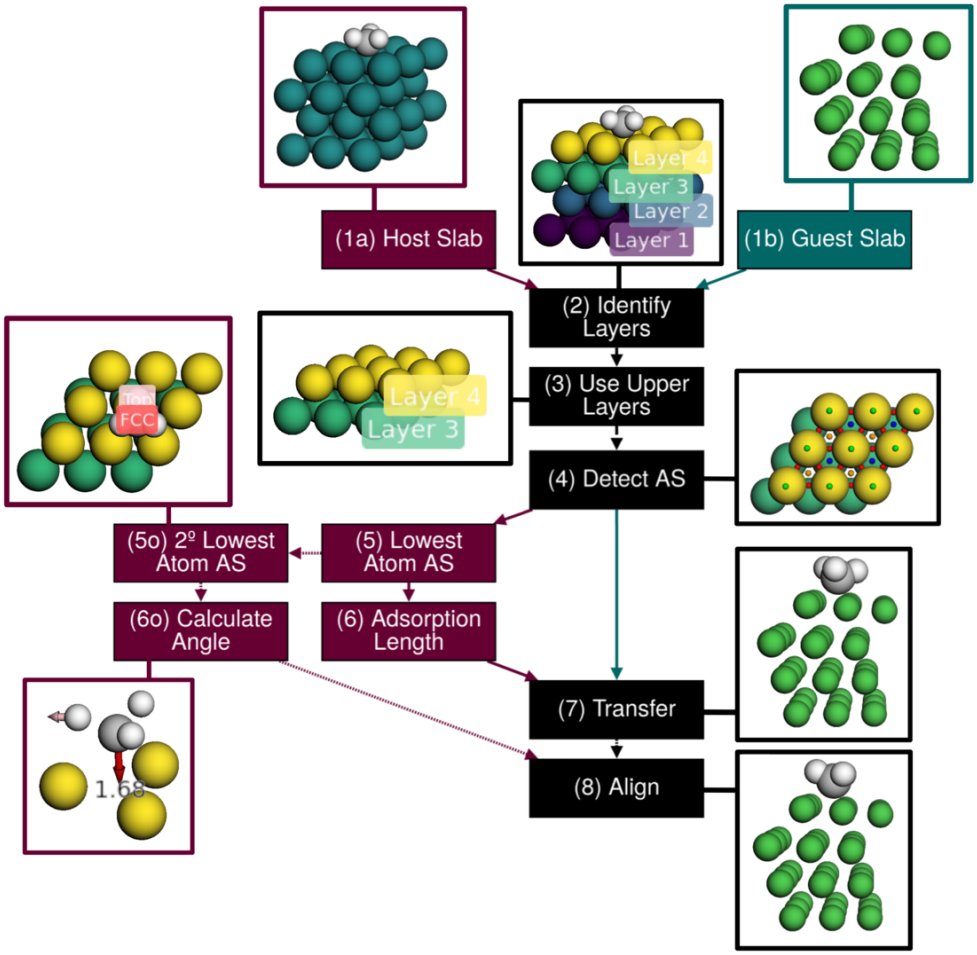

Figure 1: Interactive diagram of the steps that the transfer algorithm performs. While some of the steps only use the data from the host slab, the first (1)-(5) steps are applied concurrently to both slabs.

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, e.g. C_6 sugar alcohol decomposition consists of 10^6 intermediates (Sutton & Vlachos, 2015). 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 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 on 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 employing 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). Lastly, 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 (Isayev et al., 2017) 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

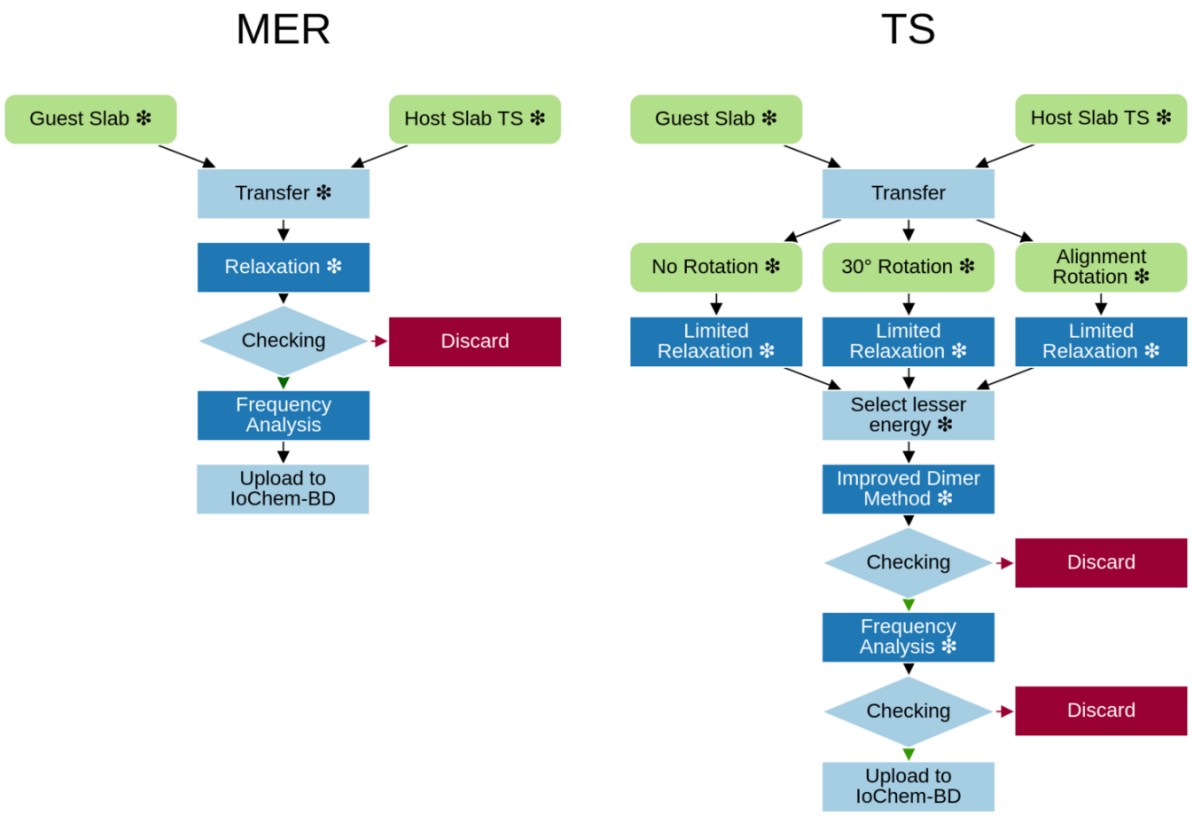

Figure 2: Interactive flux diagram showing the required steps for the relax and TS processes. DFT calculations are colored in dark blue and scripts light blue. Clicking the different points marked with () will show the status of the structure at this point.

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 minimum energy relaxation (**MER**) and another for the transition state search (**TS**). 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 relaxed geometries only generates a single geometry, the transition state (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 criteria; 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, 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. Additionally, a bond identification algorithm is used to verify that no bond breaking occurs during the relaxations.

This algorithm, that detects spurious non-valid 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 (Isayev et al., 2017) (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, an improvement has been integrated to 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. ioChem-BD (Álvarez-Moreno et al., 2014; *ioChem-BD*, n.d.) 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 an 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 is uploaded, whereas for the MER workflow the data from the relaxation method is uploaded. (Pablo-Garca, 2019; Pablo-Garca, 2020)

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 keeping 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 (f. ex. 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 to 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 contains 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 (*DCMI*, n.d.) and a 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.

Metadata	
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	

Table 1: Minimum metadata stored in ioChem-BD for each calculation.

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

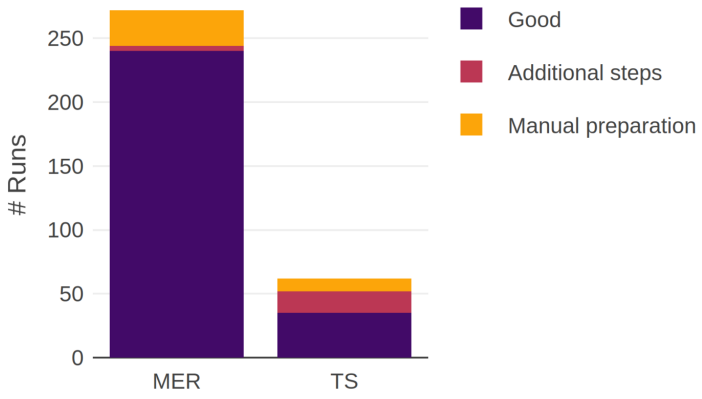

Figure 3: Efficiency of the MER and TS workflows. "Good" points to regularly terminated calculations, "Additional steps" imply that error checking routines have identified a deficiency in the calculation (i.e. maximum convergence steps reached) and "Manual preparation" points out that the calculation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

Both workflows yield reasonable results and apply the procedure for the different species (see Figure 3). Almost 88% of the minimum energy relaxations and 56% of the transition states converge at the first

attempt. Additional steps are required for 0.1% of the relaxed models and 27% of the transition states. 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, 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 transition states (TSs) there were two issues: (i) the same as for MER, insufficient ionic steps in the algorithm, this was sorted in the same manner and was by far the most common issue; (ii) 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 states end with a broken bond and more precise methods such as the Nudged-Elastic Band (NEB) (Henkelman & Jnsson, 2000; Henkelman et al., 2000) are required to obtain the geometries.

This automatization procedure has been employed to generate the databases for further machine learning studies. (Saadun et al., 2020)

3.5 Integration with Authorea

The Authorea (*Open Research Collaboration and Publishing - Authorea*, n.d.) platform allows an 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 transition state 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: (i) it enables establishing a seamless link between the computed data, the manuscript and linkes the corresponding structures interactively thus avoiding tedious and error-prone supporting information; (ii) this is particularly attractive for complex databases with massive reaction networks and/or several material compositions; (iii) the workflow reduces the computing time, systematizes the nomenclature and labeling of the different species, reduces the chaos, and increases transferability; (iv) 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; (v) the data curation is thus directly enforced by this procedure.

5 Acknowledgments

We are thankful to the "Ministerio de Ciencia e Innovacin" 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.

References

Reproducibility in density functional theory calculations of solids. (2016). *Science*, *351*(6280). https://doi.org/10.1126/science.aad3000

Fully Automated Quantum-Chemistry-Based Computation of SpinSpin-Coupled Nuclear Magnetic Resonance Spectra. (2017). *Angewandte Chemie International Edition*, *56*(46), 1476314769. https://doi.org/10.1002/anie.201708266

Machine learning in catalysis. (2018). *Nature Catalysis*, *1*(4), 230232. https://doi.org/10.1038/s41929-018-0056-y

The role of computational results databases in accelerating the discovery of catalysts. (2018). *Nature Catalysis*, *1*(11), 809810. https://doi.org/10.1038/s41929-018-0176-4

The Materials Project: A materials genome approach to accelerating materials innovation. (2013). *APL Materials*, *1*(1), 011002. https://doi.org/10.1063/1.4812323

https://materialsproject.org. https://materialsproject.org

https://nomad-coe.eu. https://nomad-coe.eu

https://www.materialscloud.org/home. https://www.materialscloud.org/home

https://cmr.fysik.dtu.dk. https://cmr.fysik.dtu.dk

The FAIR Guiding Principles for scientific data management and stewardship. (2016). *Scientific Data*, *3*(1). https://doi.org/10.1038/sdata.2016.18

Statistical learning goes beyond the d-band model providing the thermochemistry of adsorbates on transition metals. (2019). In *Nat. Comm. (In press)*.

First-principles-based multiscale modelling of heterogeneous catalysis. (2019). *Nature Catalysis*, *2*(8), 659670. https://doi.org/10.1038/s41929-019-0298-3

Machine Learning for Organic Cage Property Prediction. (2018). *Chemistry of Materials*, *31*(3), 714727. https://doi.org/10.1021/acs.chemmater.8b03572

Machine learning for molecular and materials science. (2018). *Nature*, *559*(7715), 547555. https://doi.org/10.1038/s41586-018-0337-2

Read between the Molecules: Computational Insights into Organic Semiconductors. (2018). *Journal of the American Chemical Society*, *140*(48), 1637016386. https://doi.org/10.1021/jacs.8b07985

To address surface reaction network complexity using scaling relations machine learning and DFT calculations. (2017). *Nature Communications*, *8*(1). https://doi.org/10.1038/ncomms14621

Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. (2018). *ACS Central Science*, *4*(2), 268276. https://doi.org/10.1021/acscentsci.7b00572

Machine Learning for Computational Heterogeneous Catalysis. (2019). *ChemCatChem*, *11*(16), 35813601. https://doi.org/10.1002/cctc.201900595

Machine learning meets volcano plots: computational discovery of cross-coupling catalysts. (2018). *Chemical Science*, *9*(35), 70697077. https://doi.org/10.1039/c8sc01949e

Machine Learning Accelerates the Discovery of Design Rules and Exceptions in Stable Metal–Oxo Intermediate Formation. (2019). *ACS Catalysis*, *9*(9), 82438255. https://doi.org/10.1021/acscatal.9b02165

Structure-Mechanical Stability Relations of Metal-Organic Frameworks via Machine Learning. (2019). *Matter*, *1*(1), 219234. https://doi.org/10.1016/j.matt.2019.03.002

The atomic simulation environmenta Python library for working with atoms. (2017). *Journal of Physics: Condensed Matter*, *29*(27), 273002. http://stacks.iop.org/0953-8984/29/i=27/a=273002

AiiDA: automated interactive infrastructure and database for computational science. (2016). *Computational Materials Science*, *111*, 218230. https://doi.org/https://doi.org/10.1016/j.commatsci.2015.09.013

http://openbabel.org/wiki/Main_Page. http://openbabel.org/wiki/Main_Page

Universal Chemical Synthesis and Discovery with 'The Chemputer'. (2019). *Trends in Chemistry*. https://doi.org/10.1016/j.trechm.2019.07.004

Graph Theory Approach to High-Throughput Surface Adsorption Structure Generation. (2019). *The Journal of Physical Chemistry A*, *123*(11), 22812285. https://doi.org/10.1021/acs.jpca.9b00311

Dynamic Workflows for Routine Materials Discovery in Surface Science. (2018). *Journal of Chemical Information and Modeling*, *58*(12), 23922400. https://doi.org/10.1021/acs.jcim.8b00386

A high-throughput framework for determining adsorption energies on solid surfaces. (2017). *Npj Computational Materials*, *3*(1). https://doi.org/10.1038/s41524-017-0017-z

Machine-Learning Methods Enable Exhaustive Searches for Active Bimetallic Facets and Reveal Active Site Motifs for CO2 Reduction. (2017). *ACS Catalysis*, *7*(10), 66006608. https://doi.org/10.1021/acscatal.7b01648

Accelerating the discovery of materials for clean energy in the era of smart automation. (2018). *Nature Reviews Materials*, *3*(5), 520. https://doi.org/10.1038/s41578-018-0005-z

Materials property prediction using symmetry-labeled graphs as atomic-position independent descriptors. (2019). *ArXiv e-Prints*, arXiv:1905.06048.

SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. (1988). *Journal of Chemical Information and Computer Sciences*, *28*(1), 3136. https://doi.org/10.1021/ci00057a005

FireWorks: a dynamic workflow system designed for high-throughput applications. (2015). *Concurrency and Computation: Practice and Experience*, *27*(17), 50375059. https://doi.org/10.1002/cpe.3505

Managing the computational chemistry big data problem: the ioChem-BD platform. (2014). *Journal of Chemical Information and Modeling*, *55*(1), 95103. https://doi.org/10.1021/ci500593j

https://www.iochem-bd.org. https://www.iochem-bd.org

https://authorea.com/. https://authorea.com/

Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. (1996). *Comput. Mater. Sci.*, *6*(1), 1550. https://doi.org/10.1016/0927-0256(96)00008-0

Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. (1996). *Phys. Rev. B*, *54*(16), 1116911186. https://doi.org/10.1103/PhysRevB.54.11169

Generalized Gradient Approximation Made Simple. (1996). *Phys. Rev. Lett.*, *77*(18), 38653868. https://doi.org/10.1103/PhysRevLett.77.3865

Projector augmented-wave method. (1994). *Phys. Rev. B*, *50*(24), 1795317979. https://doi.org/10.1103/PhysRevB.50.17953

Special points for Brillouin-zone integrations. (1976). *Phys. Rev. B*, *13*(12), 51885192. https://doi.org/10.1103/PhysRevB.13.5188

Efficient methods for finding transition states in chemical reactions: Comparison of improved dimer method and partitioned rational function optimization method. (2005). *The Journal of Chemical Physics*, *123*(22), 224101. https://doi.org/10.1063/1.2104507

Cytoscape: a software environment for integrated models of biomolecular interaction networks. (2003). *Genome Research*, *13*(11), 24982504. https://doi.org/10.1101/gr.1239303

3Dmol.js: molecular visualization with WebGL. (2014). *Bioinformatics*, *31*(8), 13221324. https://doi.org/10.1093/bioinformatics/btu829

Collaborative data science. (2015). Plotly Technologies Inc. https://plot.ly

Ethylene_glycol_reaction_network. (2017). Institute of Chemical Research of Catalonia. https://doi.org/10.19061/iochem-bd-1-37

https://cmr.fysik.dtu.dk/catapp/catapp.html. https://cmr.fysik.dtu.dk/catapp/catapp.html

Building large microkinetic models with first-principles accuracy at reduced computational cost. (2015). *Chemical Engineering Science*, *121*, 190199. https://doi.org/10.1016/j.ces.2014.09.011

Universal fragment descriptors for predicting properties of inorganic crystals. (2017). *Nature Communications*, *8*, 15679. https://doi.org/10.1038/ncomms15679

Automation Transfer Dataset. (2019). Institute of Chemical Research of Catalonia. https://doi.org/10.19061/iochem-bd-1-137

Full Dataset. (2020). Institute of Chemical Research of Catalonia. https://doi.org/10.19061/iochem-bd-1-150

https://www.dublincore.org. https://www.dublincore.org

Improved tangent estimate in the nudged elastic band method for finding minimum energy paths and saddle points. (2000). *The Journal of Chemical Physics*, *113*(22), 99789985. https://doi.org/10.1063/1.1323224

A climbing image nudged elastic band method for finding saddle points and minimum energy paths. (2000). *The Journal of Chemical Physics*, *113*(22), 99019904. https://doi.org/10.1063/1.1329672

Performance of Metal-Catalyzed Hydrodebromination of Dibromomethane Analyzed by Descriptors Derived from Statistical Learning. (2020). *ACS Catalysis*. https://doi.org/10.1021/acscatal.0c00679