Anonymous IJQC Reviewer[1]

[1]Affiliation not available

June 10, 2020

1 Referee Report

The paper describes a computational framework that can automate the simulations of adsorbates on a given surface. The script, starting from an initial cell containing a few layers of a substrate and an adsorbate, can detect the latter and then create a new simulation cell where the adsorbate is onto a different metal slab. Various methods are employed to determine the correct adsorbate position by automating many of the required steps.

The authors mention other relevant databases and tools existing in the field and actually use some of them. Interestingly, they show how results can be deposited in the ioChem-BD online database, facilitating access to the computational results.

The general concept is interesting, and the use of the interactive features of the Authorea platform can facilitate understanding (e.g. with Fig. 1, where one can click on the boxes describing intuitively the steps of the workflow and see the crystal structure at that specific stage, or the nice animated visualization of normal frequencies in Fig. 2).

However, I think that the manuscript has a few weaknesses that should be addressed by the authors.

- At the beginning of Sec. 3.4, authors define the results of the workflows as "solid". However, e.g. in the case of MER, 88% of the VASP relaxations just work without any need of error management, and the workflow only deals with 0.1% more. In many cases (the remaining ~23%) still there is the need of manual preparation. Therefore, this strikes me as suprising, since one of the focuses of the paper is to describe how the described platform can remove human intervention. In this case, human intervention is still needed, and it is not even significantly reduced from when VASP itself would require it (only 0.1% of the total cases, with 23% of cases where manual preparation is anyway needed). I'd like to stress that I understand that humans are helped in creating the input cells; still, the results don't strike me as "solid", and also the authors acknowledge in the abstract that performance is only "good" or even just "decent". Therefore, I don't believe that, for instance, the sentence in the conclusion "*Our framework has proven to successfully automate two different . . .*" is accurately describing the advantage of the framework over VASP itself.
- In addition to the point above, one way they mention they used to achieve convergence is to replace the metal slab with another one. But isn't this a different system? What if I really want to simulate that specific material?
- Authors mention that putting data on ioChem-BD makes their research FAIR. However, I could only find less than 10 systems in the database, while in Fig. 3 they report over 300 runs. Do they intend to make this data public to make the paper really FAIR? Otherwise, this is just a proof-of-concept demonstration but not really a FAIR paper.

- It is not clear to me what amount of reproducibility the ioChem-BD guarantees. Can the authors describe advantages and limitations of the database? For instance:

  - what is available to allow an external researcher to reproduce the simulation, and what is not?
  - Are input files of VASP available (I think some of them are available, but some only in parsed form like the initial coordinates?);
  - Are output files of VASP available (only a parsed .cml is provided? are raw outputs available? Is it possible to add a link to the CML specifications/schema? Is it possible to provide information on the code and version that performed the parsing?)
  - Is it possible to get inputs and outputs also for the other computational steps (I think only the final ones of VASP are provided?)
  - Is it possible to retrieve information on how the inputs of the simulations were obtained? (e.g. if the input of VASP was obtained by a relaxation, or the simulation was the restart of another one, is this specified somewhere?)

- Also, at the end of Sec. 3.4 they speak about NEB calculations - is is possible to inspect them and see the results?
- I believe that the paper requires an overall revision for what concerns the use of the English language.

  - There are quite a few grammar mistakes (e.g. "Our framework show" instead of "Our framework shows" in the abstract, "that can be search" instead of "that can be searched" at the end of Sec. 1, "a Gamma-centered mesh have" instead of "a Gamma-centered mesh has" in Sec. 2, "Niquel" instead of "Nickel" in one of the captions of Fig. 1, etc. (there are quite a few more occurrences later).
  - Also, the reference of "FAIR" as "functional, accessible, interoperable, recyclable" is incorrect (F is findable, R is reusable).
  - Moreover, I never found (or could find) the use of the term "avoidhuman", that also sounds to be as having a negative connotation, and I would suggest therefore to replace with some other term ("automation"?).
  - In addition, sometimes the use of wording is unusual or incorrect, and in some cases I feel that it make it hard to understand the actual meaning of some sentences. I report here some examples: "infinite xyz coordinate listing" in Sec. 1 (I guess they mean "very long" rather than infinite); the mention in the abstract that the framework performs an "experimental" procedure is very confusing (I understood much later that it is instead a computational paper, and it is not describing an experimental protocol); some sentences are long and not clear, like in Sec. 1 "As the applications grow and the access to massive computers and robust codes extends worldwide structural data, spectroscopic fingerprints, general properties can be generated as databases for molecules, nanostructures and materials." or in Sec. 2 "All the intermediates belong to the same reaction network, being the transition states all the possible elemental steps involving the intermediates.".

- At the end of Sec. 3.1, authors say "After a few tests, further improvements were integrated to the transfer algorithm." but it is not clear in detail which improvements were integrated, and the technical details of these (i.e., it is not possible with the information provided to try to reproduce their results).
- Is the code described in the paper available somewhere? In order to have a really "FAIR" and reusable dataset, it would be important to be able to rerun the same simulations/workflow.

In addition, I report here a few additional comments that the authors should probably to address to improve the overall effectiveness of the paper.

- In Fig. 2, it would be nice to be able to compare at the same time the three structures coming out of the 'transfer' step in the TS branch (now one has to do two/three clicks to switch, and cannot show

2

them at the same time, and it is therefore hard to compare them in detail)
- In Fig. 3, the use of an interactive graph is not providing any additional information or advantage. Authors might want to improve this by, e.g., providing additional information in an interactive way (e.g. by providing the materials belonging to a given part of the graph on hover, at least for the orange and red bars, or in some other way).
- Can authors describe how much the results of Fig. 3 (blue bars) are actually the results of simple VASP simulations, and how much instead possible VASP crashes are (maybe?) already partially automatically corrected thanks to the use of fireworks?
- In Fig. 2, the caption mentions squares and circles. Actually authors might want to refer to rectangles and rectangles with rounded corners, or change the shapes.
- Links to "source calculation" in the pop-ups of Fig. 2 do not work (at least for me)
- It is not clear that, in order to access the ioChemDB, one first needs to click on the "Data link" link at the beginning of 3.2, and then access to the embargoed resources is possible. It took quite a while to me to figure it out (I don't know if this is a choice of the authors, or a technical feature of Authorea).
- In the paper, the 'transition state search' is sometimes referred to as "TS", some other times as "TSS". Please make uniform.
- It is not clear what authors mean when they say, in Sec. 1, with "The process of reading documents has severely changed since the establishment of the world-wide web and the availability of more than one instance simultaneously running." - what are the instance(s) they are speaking about?
- In caption of Fig. 1, authors speak about "the first 5 steps", but it is not clear which ones they are - numbers from 1 to 5, including subletters? the top 5 boxes (numbers from 1 to 4)? Something else?
- Can authors describe in more detail how (or with which method) adsorption states are identified in step (4) of Fig. 1?
- It is interesting that it is possible to publish data on ioChem-BD with an embargo, and access for reviewers. Can authors explain what is the procedure to make the data public, at which state of publication this will happen, who is in charge of this (authors, the ioChem-BD platform maintainers, . . .) and who can guarantee that this will actually happen before publication? E.g., should the Authorea staff not make the paper public (after acceptance) until data is not made public?

Considering my comments above, I recommend that that the paper is revised by the authors before final acceptance.

2

.

2.0.1