Reviewer 1

Reviewer's comment No. 1 — [...] We therefore argued that "the expertise and time required to perform quantum-chemical calculations for atmospherically relevant molecules should constitute but a minor impediment to a wider adoption" (Wang et al., 2017). I am therefore very pleased to see that with their work, Lumiaro et al. have now obliterated even this minor impediment. While it would have been possible to make COSMOtherm-based predictions for datasets much larger than the 3414 molecules in Wang et al. (2017) using "brute force" and high-performance computing resources, Lumiaro et al. demonstrate convincingly that this can be achieved with much less computational effort using machine learning approaches.

The paper is very well written and, apart from some parts of the Methods section, easily accessible to those who are not familiar with computational chemistry and machine learning approaches.

Authors' reply: We thank the reviewer for their positive assessment of our work!

Reviewer's comment No. 2 — The compounds to which the trained algorithm was applied have very limited structural diversity (only normal decanes functionalized with up to six functional groups of only three types). Why was this relatively simple dataset of molecules generated, instead of using existing molecular datasets of atmospherically relevant species? For example, Valorso et al. (2011) generated > 200,000 oxidation products of a-pinene, i.e. one of the monoterpenes judged to be among "the most interesting molecules from a SOA-forming point of view" (line 307). A recent study generated datasets of 200,000, 550,000 and 750,000 atmospheric oxidation products of decane, toluene and a-pinene (Isaacman-VanWertz and Aumont, 2020).

Authors' reply: At the time of our study, we were not aware of the existence, or the public availability, of the datasets suggested by the reviewer. The purpose of our admittedly simple C10 dataset was not to comprehensively evaluate the performance of the algorithm (as that would in any case required extensive further COSMOtherm calculations), but just to perform a relatively simple "sanity check" of its predictions. We completely agree with the reviewer that the actual structures of the molecules in our C10 set may not be atmospherically relevant, although functional group composition certainly is.

We have now looked into the alpha-pinene dataset suggested here, but discovered that some alpha-pinene oxidation products are already included in Wang et al's dataset for which we trained our machine learning model. Testing model predictions on the same molecule class it is trained on is not good practice in ML model validation, so we did not extend our "sanity check" to these molecules. We are now building a larger dataset with an active machine learning technique and additional COSMOtherm calculations. The new dataset is based on compounds generated with the GECKO algorithm. It will be substantially larger and atmospherically more relevant than the C10 dataset. We hope to be able to report preliminary result on this work soon in a separate publication.

We clarified our motivation behind the choice of the validation dataset in the manuscript:

"While the functional group composition of our C10 dataset is atmospherically relevant, the particular molecules are not. The purpose of this dataset is to perform a relatively simple sanity check on the machine learning predictions, on a set of compounds structurally different from those in the training dataset. We note that using e.g. more atmospherically relevant compounds such as alpha-pinene oxidation products for this purpose might be counter-productive, since Wang *et al.*'s dataset used for training contains several such compounds."

Reviewer's comment No. 3 — Can the authors explain in more detail how a machine-learning model that is not fed with information on the conformations of a molecule is "capable of accounting for hydrogen-bonding interactions between functional groups" (line 366). Is this merely by structural similarity with molecules within the training set that also have such capabilities?

Authors' reply: We agree that this must be due to structural similarity in the training set. The linear structures we generate in our work do of course not have hydrogen bonds. The hydrogen bonds could therefore only be introduced by conformers. The SMILES string for all conformers of a molecule is of course the same. So if there is something in a SMILES string that indicates to the machine learning method that the structure prefers a conformer with hydrogen bonding and representative structures are in the training set, this could indeed be learned. We have clarified this in the manuscript:

"As we did not include conformational information of our C10 molecules in the machine-learning predictions, this is most likely due to structural similarities between the C10 compounds, and hydrogen-bonding molecules in the training dataset."

Reviewer's comment No. 4 — In this context, it is stated on line 380: "MBTR encoding requires knowledge of the 3-dimensional molecular structure, which raises the issue of conformer search", but section 2.2.2. does not spell out how that issue was resolved in the current study?

Authors' reply: To compute the MBTR and CM descriptors, we employed the *openbabel* software to convert the SMILES strings provided in the Wang *et al.* dataset into 3-dimensional molecular structures. Wang and collaborators must have themselves carried out a conformer search with COSMOconf, since the COSMOtherm calculations they performed typically average over many (up to 100) located conformers, but did not publish this data. Since values of KW/G, KWIOM/G and PSat were computed by averaging over conformers, there is no single conformer that correlates strongly with these values, so we decided to forgo the computationally costly conformer searches. We have now clarified this point in the manuscript:

"To compute the MBTR and CM descriptors we employed the *openbabel* software to convert the SMILES strings provided in the Wang *et al.* dataset into 3-dimensional molecular structures. We did not perform any conformer search."

Reviewer's comment No. 5 — Can the author propose how in the future, the atmospheric community will be able to obtain predictions for atmospherically relevant molecules, i.e. how a trained machine learning algorithm or its predictions could be made available for use by others. The authors still intend to improve this algorithm by extending the "training set to encompass especially atmospheric autoxidation products" (line 388), i.e. may not yet want to make the existing version accessible to others. However, it may be instructive to hear how this could look like eventually. Is it conceivable to create an easy-to-use software or webpage that is fed batches of SMILES and generates KW/G, KWIOM/G and PSat as calculated by the algorithm? Or would that take the form of a searchable database that has such algorithm-generated values stored for the "104 - 107 different organic compounds" (line 60) of atmospheric interest?

Authors' reply: Our "role model" here is the excellent and user-friendly UManSysProp webpage, where a user can insert e.g. a SMILES string, and obtain (among other things) saturation vapor pressure predictions computed using a variety of group contribution methods. We anticipate that the user interface of our model will eventually be similar to that. Ideally, in addition to providing predicted values for the different parameters, the results would also include an estimate of how reliable the predictions are (based on how similar or different the user-input molecule is to those included in the training dataset).

Reviewer's comment No. 6 — Many atmospheric applications require knowledge of phase partitioning at variable temperatures. COSMOtherm can also calculate the enthalpy of vaporization and the internal energies of phase transfer between the gas phase and water or WIOM. It would probably be advisable to eventually also train a machine learning algorithm to predict those thermodynamic properties.

Authors' reply: We agree completely. We also note, related to issues raised by the other reviewers, that predictions of various activity coefficients computable by COSMOtherm could also be useful. We changed the manuscript accordingly:

"We also intend to extend the machine learning model to predict a larger set of parameters computed by COSMOtherm, such as vaporization enthalpies, internal energies of phase transfer, and acivity coefficients in representative phases."

Reviewer's comment No. 7 — I find Figure 2 not particularly useful. While it could be beneficial to have a representation of the machine learning workflow, it should look less generic than what is depicted here. For example, "representations" make no appearance in that diagram, but are obviously an important part of the process. Also, the training and testing of the machine learning algorithm is presumably a key element of the workflow.

Authors' reply: We changed the figure following the referee's recommendation.

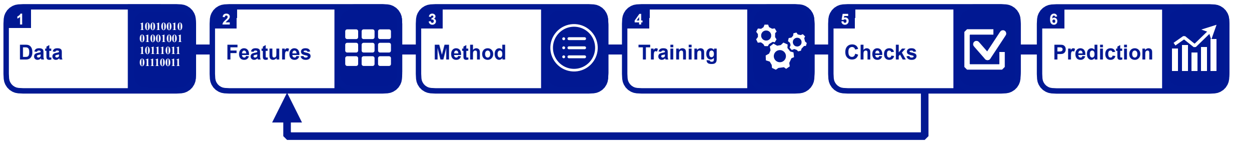

Figure 1: Schematic of our machine learning workflow: The raw input data is converted into molecular representations (referred to as features in this figure). We then set up and train a machine learning method. After evaluating its performance in step 5, we may adjust the features. Once the machine learning model is calibrated and trained, we make predictions on new data.

Reviewer's comment No. 8 — Footnote on page 2: While it is indeed quite common to estimate the KO/G by dividing KO/W by KG/W (e.g. Meylan and Howard, 2005) this is only an approximation. Whereas the octanol phase in a KO/W measurement is saturated with water and the aqueous phase is saturated with octanol, the solvents in a KW/G and KO/G measurement are typically pure. This can lead to a failure of the thermodynamic triangle to correctly estimate KO/G for hydrophobic substances (Beyer et al. 2002).

Authors' reply: Thank you for the clarification! We have changed the footnote to: "The gas-octanol partitioning coefficient ($K_{O/G}$) can then to good approximation be obtained from these by division."

Reviewer's comment No. 9 — Line 96. The abbreviation KRR is used here for the first time, but is only introduced on line 106.

Authors' reply: We removed the first instance of KRR, since it was not required on line 96.

Reviewer's comment No. 10 — Line 134: bromine not bromide

Authors' reply: Fixed

Reviewer's comment No. 11 — Line 146: The Pyzer-Knapp et al. reference is missing the year "2015" (also in the reference list)

Authors' reply: Added

Reviewer's comment No. 12 — Line 154: What does it mean if a molecular representation is "continuous"?

Authors' reply: A molecular representation is continuous, if continuous changes in the molecular structure translate into continuous changes in the representation. The many-body tensor representation (MBTR) is a good example for a continuous representation, whereas the Coulomb matrix (CM) is discontinuous. Both encode inverse distances. The MBTR does so by Gaussian broadening each inverse distance between atom pairs and then summing up these Gaussians in separate vectors for each atomic species pair. Small changes in the interatomic distances then lead to small changes in the Gaussian peak positions. Conversely, the CM assigns one value to each atom pair and collects those in a matrix whose rows and columns are sorted by their respective norm. A small interatomic distance variation could then lead to an exchange of rows and columns, which is not a continuous change of the representation.

Reviewer's comment No. 13 — Line 320: Explain the meaning of "cheaper to evaluate".

Authors' reply: The MBTR descriptor has a large data structure (22,400 vector elements) and was evaluated in several calculation stages. In contrast, TopFP is represented by a smaller data structure (8,192 vector elements) and required less computational time to evaluate, also because it did not need the conversion to 3-dimensional structures. We have now clarified in the manuscript that by ”cheaper” we refer to computational resources involved.

Reviewer’s comment No. 14 — Line 331-332: I find this sentence very confusing and I wonder whether “or less” at the end of line 331 should be deleted.

Authors’ reply: The second ”or less” was a typo, which we removed in the revised version. Thank you for spotting it!

Reviewer’s comment No. 15 — Line 336: “by almost a factor of 4000”.

Authors’ reply: “a” added as suggested

Reviewer’s comment No. 16 — Line 397 and 398: If “Zenodo, 2020” and “Gitlab, 2020” are references, they are missing from the reference list. Wouldn’t it be better to provide complete links to those datasets?

Authors’ reply: We have now updated these citations with full reference links, and DOIs where appropriate.