Atmospheric Chemistry and Physics Discussions

Atmos. Chem. Phys. Discuss., referee comment RC1
https://doi.org/10.5194/acp-2020-1258-RC1, 2021

Comment on acp-2020-1258

Frank Wania (Referee)

Referee comment on "Predicting gas–particle partitioning coefficients of atmospheric molecules with machine learning" by Emma Lumiaro et al., Atmos. Chem. Phys. Discuss., https://doi.org/10.5194/acp-2020-1258-RC1, 2021

When we first judged COSMOtherm as "the most promising approach" for predicting the equilibrium partitioning properties of SOA-related compounds (Wania et al., 2014), we qualified it as being "in the long term", because of the considerable computational effort required. Also, much less computationally costly approaches yielded predictions of the partition coefficient between a water-insoluble organic matter phase and the gas phase ($K_{WIOM/G}$) that were very similar to those obtained by COSMOtherm. However, for multifunctional compounds those simpler approaches failed to predict partition coefficients between an infinitely dilute solution in pure water and the gas phase ($K_{W/G}$) that are similar to those obtained by COSMOtherm, because only the latter accounts for the influence of conformation on intra-molecular interactions (Wang et al., 2017). 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. I have only a few questions and suggestions for improvement:

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).

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?

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?

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 $K_{W/G}$, $K_{WIOM/G}$and P_{Sat} as calculated by the algorithm? Or would that take the form of a searchable database that has such algorithm-generated values stored for the "$10^4 - 10^7$ different organic compounds" (line 60) of atmospheric interest?

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.

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.

Minor things:

Footnote on page 2: While it is indeed quite common to estimate the $K_{O/G}$ by dividing $K_{O/W}$ by $K_{G/W}$ (e.g. Meylan and Howard, 2005) this is only an approximation. Whereas the octanol phase in a $K_{O/W}$ measurement is saturated with water and the aqueous phase is saturated with octanol, the solvents in a $K_{W/G}$ and $K_{O/G}$ measurement are typically pure. This can lead to a failure of the thermodynamic triangle to correctly estimate $K_{O/G}$ for hydrophobic substances (Beyer et al. 2002).

Line 96. The abbreviation KRR is used here for the first time, but is only introduced on line 106.

Line 134: bromine not bromide

Line 146: The Pyzer-Knapp et al. reference is missing the year "2015" (also in the reference list)

Line 154: What does it mean if a molecular representation is "continuous"?

Line 320: Explain the meaning of "cheaper to evaluate".

Line 331-332: I find this sentence very confusing and I wonder whether "or less" at the end of line 331 should be deleted.

Line 336: “by almost a factor of 4000”.

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?

References

Beyer, A., Wania, F., Gouin, T., Mackay, D., and Matthies, M.: Selecting internally consistent physicochemical properties of organic compounds. Environ. Toxicol. Chem. 21, 941-953, 2002.

Isaacman-VanWertz, G., and Aumont, B. Impact of structure on the estimation of atmospherically relevant physicochemical parameters. https://doi.org/10.5194/acp-2020-1038

Meylan, W.M. and Howard, P. H.: Estimating octanol–air partition coefficients with octanol–water partition coefficients and Henry’s law constants. Chemosphere, 61, 5, 640-644, 2005.

Valorso, R., Aumont, B., Camredon, M., Raventos-Duran, T., Mouchel-Vallon, C., Ng, N. L., Seinfeld, J. H., Lee-Taylor, J., and Madronich, S.: Explicit modelling of SOA formation from a-pinene photooxidation: sensitivity to vapour pressure estimation, Atmospheric Chem. Phys., 11, 6895–6910, 2011.

Wang, C. Yuan, T., Wood, S. A., Goss, K.-U., Li, J., Ying, Q. and Wania, F.: Uncertain Henry’s law constants compromise equilibrium partitioning calculations of atmospheric oxidation products. Atmos. Chem. Phys., 17, 7529-7540, 2017.

Wania, F., Lei, Y. D., Wang, C., Abbatt, J. P. D., and Goss, K.U.: Novel methods for predicting gas-particle partitioning during the formation of secondary organic aerosol. Atmos. Chem. Phys. 14, 13189–13204, 2014.