Atmospheric Chemistry and Physics Discussions

Atmos. Chem. Phys. Discuss., referee comment RC2
https://doi.org/10.5194/acp-2020-1258-RC2, 2021

Comment on acp-2020-1258

Anonymous Referee #2

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-RC2, 2021

Review of Predicting Gas-Particle Partitioning Coefficients of Atmospheric Molecules with Machine Learning by Lumiaro et al.

It was interesting to read this manuscript. The topic of the manuscript is the prediction of saturation vapor pressures and partitioning coefficients between the gas phase and an aqueous phase and an organic phase respectively relevant in atmospheric science. There is a lack of experimental data on such properties and given the overwhelming amount of different molecules in the atmosphere, reliable computational methods that can predict such properties for a large amount of molecules are valuable. In this work, the authors explore the use of a machine learning method to predict selected thermodynamic properties for a large number of molecules, which seems very promising and timely.

Having said this, I also need to say, that I find the manuscript difficult to read in some aspects, and that the link to atmospheric chemistry and physics could be better explained and discussed. I am not an expert on machine learning and so many of my suggestions below are suggestions for improving the manuscript in relation to atmospheric relevance, which I hope the authors will find useful. To improve the manuscript I also find that the authors should give a clearer description of the thermodynamic framework for what they are calculating and discuss the limitations and relevance.

Major comments

References – I do not find that there are enough references to the literature throughout the introduction. As an example statements like "They scatter and absorb solar radiation and form cloud droplets in the atmosphere, affect visibility and human health and are responsible for large uncertainties in the study of climate change." and "Most aerosol particles are secondary organic aerosols" should be accompanied by one or more literature references. Likewise, in section 4 on prediction I miss examples and references for the

statements for example on functionalization and fragmentation.

The thermodynamic basis – vapor pressures and partitioning coefficients.

I expect several of the low volatile species will be solids at room temperature and likely exist in the subcooled liquid state in the atmosphere. There can be a large difference between the vapor pressure of the solid and that of the subcooled liquid. I assume the vapor pressures calculated are for the subcooled liquid state. This should be specified. Likewise, it should be better explained to the reader what the physical meaning of the partitioning coefficients is? Do they represent partitioning over a flat surface? It says they are infinite dilutions – does this mean the activity coefficients are one? What values are assumed for the activity coefficients? partitioning in the atmosphere depends on many things including particle size, amount of condensed material, accommodation coefficients – I suggest this is recognized and addressed.

Where does the formula for calculation of saturation vapor pressure come from? Please give a derivation or a reference. The saturation vapor pressure is a property of the pure component – but here it seems to depend on the activity in a mixture and a partitioning coefficient? The equilibrium vapor pressure over a mixture depends on the activity?

What is meant with the statement "Saturation vapor pressure describes the interaction of a compound with itself" (page 2 line 29/30) ? and "partitioning coefficients (K) for the interaction of the compound with representative other species." I would say, that it is the activity coefficients that account for interactions between molecules in the condensed phase. In the gas phase – do the authors consider molecular interactions?

General comments

Some sentences are unclear: eg. "For relatively simple organic compounds, efficient empirical parametrizations have been developed to predict their condensation-relevant properties. " – the authors should help the reader here with more clear definitions - what is a "relatively simple organic compound" – and what are the exact condensation relevant properties and which efficient empirical parameterizations are the authors referring to here (references should be given) ?

To help the reader I also suggest to restructure the manuscript a bit and define the coefficients that are modelled already in the introduction.

How was vapor pressures obtained/calculated from COSMOtherm – this is unclear from the manuscript and should be specified.

Could the authors reflect on why the MBTR method performs so much better than the other methods?

Accuracy and performance: It should be stated explicitly what the COSMOTHERM accuracy is, both on the predicted saturation vapor pressures and on the partitioning coefficients.

Page 7 line 158 – what is "good performance" ?

Atmospheric context

I miss a short description of which parent VOCs were considered for the basis set used.

Regarding the prediction section. As the authors write monoterpenes are relevant molecules and as I understand the choice of 10 carbon atoms is based on monoterpenes. The choice of a linear alkane chain is motivated by simplicity – but is it relevant in the atmosphere from monoterpene oxidation? Are all the molecules studied in the master chemical mechanism? – I would have expected at least some molecules with a ring structure included.

The authors several times discuss formation of particles and – is there a reference for some thought of threshold vapor pressure value ? For example Page 2 line 50 a threshold value of 10-12 Pa for nucleation is given.

In the abstract it says" The resulting saturation vapor pressure and partitioning coefficient distributions were physico-chemically reasonable, and the volatility predictions for the most highly oxidized compounds were in qualitative agreement with experimentally inferred volatilities of atmospheric oxidation products with similar elemental composition."

I do not see justification for this in the manuscript. I miss examples (optimally for all the compounds) where the authors give the experimental vapor pressure, the vapor pressure obtained from a state of the art group contribution method, the cosmotherm vapor pressure and the vapor pressure obtained using the machine learning code and discuss differences and similarities. For the lowest vapor pressures experimental data are not available. The authors should give the range of vapor pressures where the model can be compared with experimental data. It is not clear what is meant with elemental composition – normally the molecular formula or even structural formula is needed to predict a vapor pressure?

Other

Page 2 line 3: Several experimental techniques are capable of measuring saturation vapor pressures of 10-5 Pa. It would be appropriate to cite literature providing experimental vapor pressures. What is the definition of non-volatile that the authors use?

Page 3 line 63: "Here, we take a different approach compared to previous parametrization studies, and consider a data-science perspective (Himanen et al., 2019). Instead of assuming chemical or physical relations, we let the data speak for itself." - what is meant with letting the data speak for itself?

Figure 9 b: what is on the y-axis - is it a percentage? or an absolute number?

Page 16: "This result demonstrates that unlike the simplest group-contribution models (which would invariably predict that the lowest-volatility compounds in our C10 dataset should be the tetrahydroxydicarboxylic acids), both the original COSMOTherm predictions, and the machine-learning model based on them, are capable of accounting for hydrogen-bonding interactions between functional groups."

I am not sure this statement is quite fair -- to my knowledge state of the art group contribution methods (e.g. those on the UMAN Sysprop webpage) include interactions – which simple group contribution methods are the authors referring to and are such simple methods being used in atmospheric simulations?