Atmospheric Chemistry and Physics Discussions

Atmos. Chem. Phys. Discuss., referee comment RC3
https://doi.org/10.5194/acp-2020-1258-RC3, 2021

Comment on acp-2020-1258

Anonymous Referee #3

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-RC3, 2021

The authors utilize machine learning to predict saturation vapor pressure and two equilibrium-partitioning coefficients for gas-particle partitioning. For training and validating the machine learning model they use a dataset obtained by COSMOtherm calculations of theses observables for atmospheric oxidation product molecules.

The paper is well written, the topic timely and of great interest for the readers of ACP and I recommend publishing but ask the authors to take the following comments and suggestions into account.

I have one very general concern, which does not relate to the machine learning approach presented here, but to the underlying COSMOtherm data set. The authors write (e.g. line 49 page 2) that the COSMOtherm predictions have an order of magnitude accuracy. However, for a number of compounds at low saturation vapor pressures there have been studies comparing experimental saturation vapor pressures with COSMOtherm predictions and finding much larger deviations (e.g. Bannan et al., 2017, Krieger et al. 2018). It should be pointed out that the COSMOtherm model has been "calibrated" with a parametrization dataset of known compounds, which are potentially biased to high saturation vapor pressures (Klamt et al. 1998). Therefore, the accuracy of the underlying reference data may be only several orders of magnitude for low saturation vapor pressure components.

For gas-particle partitioning, the saturation vapor pressure range from about 10^{-11} kPa to about 10^{-3} kPa is relevant (e.g. Valorso et al. 2011, or the discussion starting in the last paragraph of page 2). However, Fig. 3c shows that there are hardly any molecules in the dataset below 10^{-8} kPa. Actually about half of the dataset contains compounds, which will be entirely in the gas phase under atmospheric conditions. Does this pose a problem?

Related: the last paragraph on page 6 states that Wang's dataset is rather small for

machine learning but internally consistent. I intuitively understand that this helps the machine-learning model to succeed in predicting well. However, the authors write that Sanders's dataset for 17350 Henry's law constant are not internally consistent (as Wang's dataset). But what if the Sander's data are the correct ones? What if the real world is more complex than what is predicted by COSMOtherm? Would the machine learning approaches fail because it there are no easy "rules" the machine-learning algorithm can pick out of the dataset? Would the output of a model trained with these data just produce random partitioning coefficients within the range of the data set? These questions are probably impossible to answer without doing the experiment. It would have been very interesting to see how the machine-learning model perform on the dataset of Sander, but this is clearly beyond the work presented here.

I find section 2.2.4 rather brief. For me – being not familiar with the topic – it is not possible to follow despite Fig. 4d. May be extent a bit?

Discussion on page 16: Related to my comments above, without experimental vapor pressures for the C10 compounds being available, this discussion is interesting, but there may be surprises if experimental vapor pressures become available. I feel the authors should clearly state that the COMOtherm predictions are not validated in this pressure regime at all.

Technical comment:

Page 12, line 292: Figure 5 should be Fig. 3, correct?

References:

Bannan, T. J. et al.: Measured Saturation Vapor Pressures of Phenolic and Nitro-aromatic Compounds, Environ. Sci. Technol. 2017, 51, 3922−3928.

Klamt, A.; et al.: Refinement and Parametrization of COSMO-RS, J. Phys. Chem. A 1998, 102, 5074-5085.

Krieger , U. K. et al.: A reference data set for validating vapor pressure measurement techniques: homologous series of polyethylene glycols, Atmos. Meas. Tech., 11, 49–63, 2018.

Valorso, R. et al.: Explicit modelling of SOA formation from -pinene photooxidation: sensitivity to vapour pressure estimation, Atmos. Chem. Phys., 11, 6895–6910, 2011.