Responses to Reviewers: Ch3MS-RF

Reviewer 2

This study developed new machine learning techniques to characterize unidentifiable organic compounds using GC-MS and GCxGC-MS techniques. The authors provided a detailed discussion and demonstration of this model and its potential to improve the current understanding of undefined organic species in the atmosphere. This new method is able to improve the quantification accuracy compare with manual proxy modeling, which will lead to a better understanding of atmospheric organic aerosols formation and chemical properties. I'm supportive of this paper and recommend for publication in Atmospheric Measurement Techniques.

Here are a few minor comments:

Authors have mentioned that the vapor pressures were calculated for model training and evaluation, a few external standards test set and extrapolation set species were incompatible with vapor pressure prediction, can the author provide more explain more about how it is incompatible and the evaluation process for vapor pressure? Based on Figure 4, it seems like the predicted vapor pressure has more variability than other perimeters, and the more accurate vapor pressure can improve the model accuracy.

We thank reviewer 2 for this opportunity to clarify the reasoning behind restricting the vapor pressure training and prediction datasets. The Nanoolal and SIMPOL vapor pressure estimates were generated using the GECKO-A tool, which in its current form does not allow certain molecular structures (eg PAHs) or functional groups (e.g. amines) to be processed. While some of these species could have been predicted by hand, manual vapor pressure predictions were not feasible at the number scale of compounds utilized in this analysis. The referenced version of EVAPORATION does not contain parameterizations for some functional groups such as amines and heterocyclic compounds, rendering these species outside the scope of estimates using these methods. Under ideal circumstances, validated experimental vapor pressures for all test and training compounds would be used to produce the most accurate predictions possible, but experimental vapor pressures are not available for many more species than were excluded due to lack of structurally predicted vapor pressures and other compounds have conflicting experimental vapor pressures reported in the literature. The method described in this work was selected for a few reasons. First, there is a high overlap between compounds whose vapor pressures could be predicted by SIMPOL, EVAPORATION, and Nannoolal and compounds in our training test and validation sets. Second, the referenced work in Isaacman-Van Wertz and Aumont et al., 2021 finds that the average of these methods produces optimally accurate vapor pressure predictions for the compound classes most commonly observed in the extrapolation set. Third, the structural vapor pressure predictions were both efficient and internally consistent, particularly when compared to compiling often conflicting values from the literature. The following language has been added (line 318) to clarify the limitations which rendered some compounds incompatible with predictions.

"Seven of the external standard test set species and fifteen of the extrapolation set species were incompatible with the prediction capabilities of one or more of the three structural vapor pressure

prediction methods (most frequently due to functional group types for which the models are not parameterized) and were therefore not utilized in performance analysis."

Author mentioned that the model underestimated the high carbon oxidation state region and the high carbon number region, but there is no predicted data shown in the plot? it also seems like the model is a little bit overpredicted for the carbon number region between 20 and 30, can author comment on that?

Thank you so much for this very helpful note- we intended to state that the high carbon oxidation state compounds (which have low carbon numbers) have underpredicted carbon oxidation states, and that the high carbon number compounds (which have low carbon oxidation states) have underpredicted carbon oxidation states- however this was unclear and we see how this may have been interpreted as referring to compounds that have both high carbon number and high carbon oxidation state, of which there were none. The wording has now been altered (line 498) to clarify that these are independent observations and not descriptive of any group of compounds.

> "This is apparent in both the high $\overline{OS_c}$ region and the high carbon number regions of the $\overline{OS_c}$-n_c space, where high carbon oxidation states and high carbon numbers were each independently underpredicted."

For the compounds in the region between carbon number 20 and 30, while the distribution of predicted values in this region is higher than observed, this effect is primarily caused by the underprediction of carbon numbers, as now clarified in the description of Figure 5. This is to say, the predicted points are shifted to the right of the true properties positions for those compounds rather than shifted up. We believe that the clarification regarding underprediction of carbon numbers for high carbon number species now clarifies this point.

> "As illustrated in Figure 5, the real and predicted chemical properties spaces for the ambient data set indicate both strengths and weaknesses for this application of chemically properties modelling. As noted earlier, random forest modelling does not extrapolate and has a tendency to underpredict property extremes. This is apparent in both the high $\overline{OS_c}$ region and the high carbon number regions of the $\overline{OS_c}$-n_c space, where high carbon oxidation states and high carbon numbers were each independently underpredicted. These errors could be moderated by adding more oxygenated species and higher carbon number species to the external standard, which would provide the model with more information to predict properties in these regions."

Does this model capable of any GC-MS system or is there any specific requirement for the instrument? Can author add some discussion of the limitation of this model as well?

Thank you for this comment, which is highly aligned with similar questions from Reviewer 1. To better address appropriate considerations in adapting this technique to different instruments and for incorporating data produced by multiple instruments and techniques, we have now added section 5.3 *Considerations for Adaptation Across Instruments and Methods* and figure 8, reproduced below.

5.3 "Considerations for Adaptation Across Instruments and Methods

The approach presented in this work prioritizes continuity between training, test, and sample data by exclusively training the model on data produced by a single instrument using a standardized methodology. This approach was selected to ensure that the patterns identified by Ch3MS-RF modelling in the training data were as directly relevant as possible to the unidentifiable sample compounds of interest. However, in some cases, accumulation of a representative external standard spanning the entire feature domain of unidentifiable compounds of interest may not be practical or possible. Electron ionization (70 eV) mass spectrometry is an extremely well characterized and consistent technique, but chromatographic retention times and indices can vary. In order for data produced by multiple instruments and techniques to be integrated within Ch3MS-RF, it is therefore important to establish the tolerance of prediction performance to drifts in retention index.

To test sensitivity to retention index or retention time shifts across instruments and methods, the vapor pressure, carbon number, $\overline{OS_c}$, and O:C of the external standard test set compounds were predicted using retention index inputs that were shifted from their observed retention indices. A broad range of shifts from -200 (indicating the equivalent of a two-carbon number shift, for example if in the test sample heptadecane were to elute at the time that pentadecane eluted in the training standard run) to +200 were tested (including -200, -150, -100, -50, -25, +25, +50, +100, +150, +200). A new mean average error was calculated for each set of predictions based on the shifted retention indices and compared to the unshifted mean average error to calculate the % increase in mean average error as a function of test set retention index shift. These results are visualized in Figure 8. The two measures of oxidation, $\overline{OS_c}$ and O:C were relatively insensitive to retention index shifts, as their mean average errors increased by less than 10% at a retention index shift of ±200 and by < 5% within retention index shifts of ±100. Carbon number and vapor pressure predictions were more sensitive to retention index shifts, as would be expected given that retention times are more directly physically related to these two properties. At retention index shifts of + 200, mean average error of carbon number prediction increased by 44%, while a shift of -200 produced vapor pressure predictions that increased by 39%, both of which significantly decrease the utility of the produced predictions. However, within retention index shifts of ±100, increases in vapor pressure and carbon number prediction errors are modest, with all calculated MAE % error increases < 10%, with the exception of a 12% increase in error for vapor pressure predictions at a retention index shift of -100. Vapor pressure prediction in fact appears to slightly improve at shifts of +<25-50, but these improvements are extremely modest (<3%), are attributable to the generally higher uncertainties in vapor pressure prediction, and are not significantly different from predictions produced at a retention index shift of 0. Reported n-alkane normalized kovats indices of compounds within standardized column types (semistandard non-polar, standard non-polar, etc.) typically vary by <50, meaning that where methodologies allow test compound kovats or retention indices to be calculated, predictions utilizing training data from instruments and analysis protocols not used on the test samples are likely to be robust, particularly for O:C and $\overline{OS_c}$. For methodologies that do not use internal standards and that cannot otherwise easily yield kovats indices, protocols using similar columns and temperature ramps would likely produce retention times that could be substituted for retention indices in the Ch3MS-RF methodology. This approach would be usable across multiple instrumentations, provided it could be established that

the retention times of any given compound produced by the training and test instrument drift by less than 1 carbon number equivalent.

In summary, training and/or test data from multiple instruments and protocols can be combined to meet user needs, provided the following criteria are met: 1) the same ionization energy (typically 70 eV) is used 2) retention index or retention time drifts between instruments or protocols can be normalized to less than the difference of the elution time between two sequential linear alkanes (retention index drift of <100) 3) similar phase columns are used (semistandard nonpolar, standard nonpolar, etc) 4) samples and training data are consistently either derivatized or underivatized, and if derivatized use a consistent derivatization agent. It is also important to keep in mind that the training data must span the anticipated feature space of the use data set, and that in cases of doubt this can be tested by adding extrapolation set compounds identified from the sample medium. For chemical properties modelling, this approach can be adapted from the GCxGC approach presented for any instrument using chromatography- electron ionization-mass spectrometry that has the capacity to yield at least unit resolution mass spectra and for which spectra can be sufficiently deconvoluted to yield clean analyte spectra. The model structure and provided sample code are highly flexible and could be utilized to predict any property of interest that might reasonably be expected to be reflected in the combination of compound mass spectra and chromatographic retention time, although performance evaluation is always important for ensuring that the patterns are sufficiently strong to enable accurate property prediction using Ch3MS-RF.
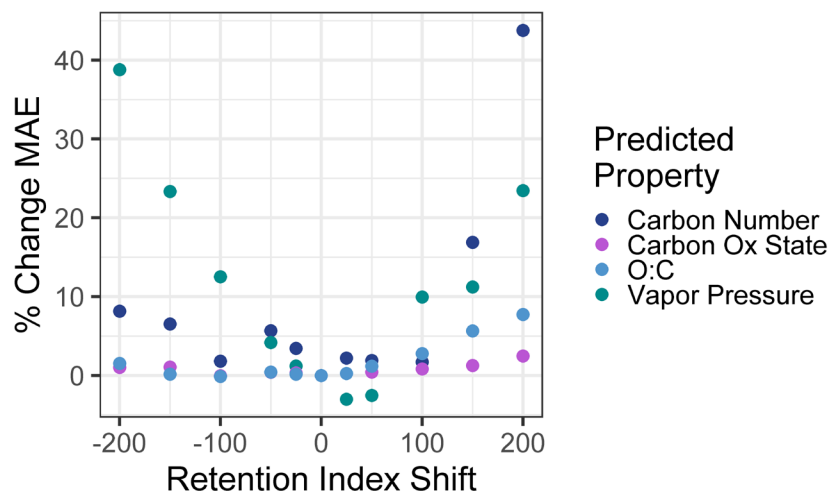

Figure 8: % increases of mean average error in chemical property prediction as a function of shift in test set retention index relative to training set retention index. Retention indices are normalized to a linear alkane series, making an increment of 100 indicate the retention time differences between two linear alkanes separated by 1 carbon number.