Geosci. Model Dev. Discuss.,
doi:10.5194/gmd-2016-312-RC1, 2017

Geoscientific Model Development Discussions

Interactive comment on "STRAPS v1.0: Evaluating a methodology for predicting electron impact ionisation mass spectra for the aerosol mass spectrometer" *by* David O. Topping et al.

Anonymous Referee #1

Received and published: 17 February 2017

In this work, the authors harvest a set of molecular descriptors from various molecules and establish relationships to mass fragments observed by aerosol mass spectrometry. They evaluate different sets of molecular descriptors and supervised learning methods to evaluate the range of predictive capability that can be achieved for a set of individual compounds, and also for mixtures derived from chemically explicit simulation. While inherent differences between model simulations and reality preclude strict comparisons, they also evaluate the general trajectory in the evolution of the predicted f44 to simulated O:C ratio and f44 to f43. The idea presented in this manuscript is a nice one. A link between chemical composition and AMS mass spectra would be desirable; the challenges for predicting electron impact mass spectra from first principles and justifi-

cations for taking a chemoinformatic/statistical approach are outlined well. Statistical models relating molecular properties to kinetic rate constants (structure-activity relationships) are widely accepted in the community (e.g., Carl et al., 2007), so an effort such as this one relating molecular properties to observable instrumental signals are a welcome addition. The manuscript is recommended for publication by Geoscientific Model Development; addressing the following comments may improve the readability of the manuscript.

The authors highlight the "proof-of-concept" nature of the study with many issues to be resolved in future work, which is understandable given its novelty. However, the main achievements in this work are not highlighted well. Is the MACCS fingerprints most successful just because of the sheer number of keys, each of which contribute to predictions, or are there particular structural elements not present in the others that improve the predictions? The generally poor performance of SVMs for all keys is surprising, is it possibly due to the high dimensionality in the underlying representations that is not present in the others, or is there a more obvious reason to the authors?

How are the tuning parameters for the model parameters determined? For instance, the penalty factor for SVM, etc.?

Are cosine angles (uncentered correlations) sufficient to capture agreement that represents more than the range (minimum and maximum) relative ion counts for each spectrum? This angle may not represent disagreement in relative ion counts that are of intermediate value very well. In that there is precedent for cosine angles for mass spectra comparison, it is a safe metric, but the authors may look at analyzing residuals for each mass fragment to understand what their model gets right and less right (to generalize on illustrations provided in Figures 5 and 6, which are incidentally missing axes labels). There is some mention about f43 being somewhat reasonable and f44 being underpredicted, but this seems a bit buried in the presentation.

Is the number of keys used vs. m/z variables and issue? Given the smaller number

of non-zero keys and number of samples, is it reasonable to try to predict 300 m/z's in the AMS spectrum (In Figures 5-8 only 100 are shown, but is the model trained only to predict 100 m/z's)? Would not the authors benefit from trying to reproduce a "reduced" set of spectra (e.g., reconstructed from a truncated set of PCA or PMF components)? Is there a reason why all keys were not combined into a single fingerprint? It would be simple to remove redundant keys simply by inspection, if that were a concern.

Regarding the comparison of f44 and O:C (Figure 8), is not the COO^+ associated with m/z 44 more sensitive to dicarboxylic acids (Russell et al., 2009)?

A minor point: The simulation (photoxidation) conditions of Valorso (2011) can be repeated in the caption of Figure 9 so the reader can immediately contextualize the comparison.

References:

S. A. Carl, L. Vereecken, and J. Peeters, Phys. Chem. Chem. Phys., 2007, 9, 4071-4084 doi: 10.1039/B705505F.

L. M. Russell, R. Bahadur, L. N. Hawkins, J. Allan, D. Baumgardner, P. K. Quinn, T. S. Bates, Organic aerosol characterization by complementary measurements of chemical bonds and molecular fragments, 2009, 43, 6100–6105, doi:10.1016/j.atmosenv.2009.09.036.

Interactive comment on Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-312, 2017.