STRAPS v1.0: Evaluating a methodology for predicting electron impact ionisation mass spectra for the aerosol mass spectrometer

David .O. Topping[1,2], James. Allan[1,2], M. Rami, Alfarra[1,2], and Bernard Aumont[3]

[1]School of Earth and Environmental Science, University of Manchester, Manchester, M13 9PL, United Kingdom
[2]National Centre for Atmospheric Science, University of Manchester, Manchester, M13 9PL, United Kingdom
[3]LISA, UMR CNRS 7583, Universite Paris Est Creteil et Universite Paris Diderot, Creteil, France

Correspondence to: David. O. Topping (david.topping@manchester.ac.uk)

Abstract. Our ability to model the chemical and thermodynamic processes that lead to secondary organic aerosol (SOA) formation is thought to be hampered by the complexity of the system. While there are fundamental models now available that can simulate the tens of thousands of reactions thought to take place, validation against experiments is highly challenging. Techniques capable of identifying individual molecules such as chromatography are generally only capable of quantifying a subset of the material present, making it unsuitable for a carbon budget analysis. Integrative analytical methods such as the Aerosol Mass Spectrometer (AMS) are capable of quantifying all mass, but because of their inability to isolate individual molecules, comparisons have been limited to simple data products such as total organic mass and O:C ratio. More detailed comparisons could be made if more of the mass spectral information could be used, but because a discrete inversion of AMS data is not possible, this activity requires a system of predicting mass spectra based on molecular composition.

In this proof of concept study, the ability to train supervised methods to predict electron impact ionisation (EI) mass spectra for the AMS is evaluated. Supervised Training Regression for the Arbitrary Prediction of Spectra (STRAPS), is not built from first principles. A methodology is constructed whereby the presence of specific mass-to-charge ratio (m/z) channels are fit as a function of molecular structure before the relative peak height for each channel is similarly fit using a range of regression methods. The widely-used AMS mass spectral database is used as a basis for this, using unit mass resolution spectra of laboratory standards.

Key to the fitting process is choice of structural information, or molecular fingerprint. Our approach relies on using supervised methods to automatically optimise the relationship between spectral characteristics and these molecular fingerprints. Therefore, any internal mechanisms or instrument features impacting on fragmentation are implicitly accounted for in the fitted model. Whilst one might expect a collection of keys specifically designed according to EI fragmentation principles to offer a robust basis, the suitability of a range of commonly available fingerprints is evaluated.

Using available fingerprints in isolation, initial results suggest the generic public 'MACCS' fingerprints provide the most accurate trained model when combined with both decision trees and random forests with median cosine angles of 0.94-.0.97 between modelled and measured spectra. There is some sensitivity to choice of fingerprint, but most sensitivity is in choice of regression technique. Support Vector Machines perform the worst, with median values of 0.78-0.85 and lower ranges approaching 0.4 depending on the fingerprint used. More detailed analysis of modelled versus mass spectra demonstrates important composition dependent sensitivities on a compound-by-compound basis. This is further demonstrated when we apply the trained methods to a model α-pinene SOA system, using output from the GECKO-A model. This shows that use of a generic fingerprint referred to as 'FP4' and one designed for vapour pressure predictions ('Nanoolal') give plausible mass spectra, whilst the use of the MACCS keys in isolation perform poorly in this application, demonstrating the need for evaluating model performance against other SOA systems rather than existing laboratory databases on single compounds.

Given the limited number of compounds used within the AMS training dataset, it is difficult to prescribe which combination of approach would lead to a robust generic model across all expected compositions. Nonetheless, the study demonstrates the use of a methodology that would be improved with more training data, fingerprints designed explicitly for fragmentation mechanisms occurring within the AMS, and data from additional mixed systems for further validation. To facilitate further development of the method, including application to other instruments, the model code for re-training is provided via a public Github and Zenodo software repository.

1 Introduction

Volatile organic compounds (VOCs), emitted from both natural and anthropogenic sources, are oxidised in the atmosphere to form lower-volatility species that condense onto aerosol particles or contribute to new particle formation (Laaksonen et al., 2008;Sipila et al., 2016;Ehn et al., 2014). With an enormous number of species that are present, this diversity in chemistry is reflected in the extensive range of species and chemical signatures identified in ambient studies (Hamilton et al., 2013). Within atmospheric science, it is desirable to develop models for secondary organic aerosol (SOA) formation based on a given set of precursors and photochemical processing. Within most global and regional models, often-used techniques include modelling representative photochemical yields from specific precursors and tuning accordingly (Spracklen et al., 2011) or employing a parametric model such as the volatility basis set (Robinson et al., 2007). While both of these approaches can deliver realistic absolute concentrations, because they are not based on explicit physical processes, their predictive skill is always subject to question (Hallquist et al., 2009;Bergstrom et al., 2012). It is therefore desirable to develop SOA models based around actual molecular processes and kinetics constrained through laboratory experiments (where available), such that this skill can be evaluated. Such models rely on explicit chemical mechanisms such as the Master Chemical Mechanism (MCM) (Saunders et al., 1997) or the GECKO model (Aumont et al., 2005). While this mechanistic approach has resulted in poor performance in terms of absolute mass concentrations in the past (Volkamer et al.,

2006), much of this shortfall can be accounted for by not considering all precursors (in particular the semi-volatile and intermediate-volatility organic matter), unexpected processes likely to produce lower-volatility products (e.g. oligomerisation and autoxidation (Ehn et al., 2014) and inadequacies associated with phase partioning models (Barley and McFiggans, 2010;Valorso et al., 2011;McVay et al., 2016). As the availability of data regarding these has improved and thus our understanding of these processes matured, the performance of the models has become more realistic (McVay et al., 2016). The development of more applicable explicit models has been facilitated by the ability to automatically predict processes rather than prescribe them (Aumont et al., 2012;Aumont et al., 2005) as has been implemented in the Generator of Explicit Chemistry and Kinetics of Organics in the Atmosphere (GECKO-A) and the forthcoming version 4 of the MCM (http://gotw.nerc.ac.uk/list_full.asp?pcode=NE%2FM013448%2F1). This can be supplemented by the automated prediction of properties important for partitioning, using generalised informatics tools such as UManSysProp (Topping et al., 2016). While it is unlikely that such complex models would be used directly for large-scale Eularian chemical transport and climate models, and uncertainties with regards to fundamental properties remain (Bilde et al., 2015), they are still highly useful for benchmarking and providing the parameters for simpler models.

Comparison of model output with measurements in the ambient air and in the laboratory is required to test model accuracy. With current analytical methods, it is impossible to detect and quantify every compound in the particle even if we can predict compound-by-compound speciation. While there are techniques capable of resolving a large number of molecules such as electrospray ionisation and two-dimensional gas chromatography (Noziere et al., 2015), comprehensively calibrating for and thus providing quantitative data on the abundances of the molecules is difficult. The AMS, which is often used in chamber and flow tube experiments, is capable of delivering data on the total mass concentration of organic matter and some other simple top-down metrics such as the O:C ratio (Aiken et al., 2007). However this does not provide the ideal constraint of such models.

While the mass spectral data can be further investigated through inspection of markers at specific m/z channels (such as 43 and 44) (Ng et al., 2011), such data tends to be qualitative and result in speculative conclusions (Morgan et al., 2010). In theory, the data across the mass spectrum could be more systematically compared with the modelled data if knowledge of the instrument response to molecular features could be invoked in a general fashion (Ehn et al., 2014).

In this proof of concept study we evaluate a methodology to bridge existing model measurement comparison. A database of the AMS mass spectral responses to various molecules has been built up over the years and this has been used to characterise the response of certain key peaks to certain functional groups (Ulbrich et al., 2009;Ehn et al., 2014). In this study we use that information to develop and evaluate regression software that predicts an AMS spectrum based on the predicted aerosol composition (figure 1)

This is not the first study on predicting EI mass spectra based on molecular composition, or to demonstrate the potential for predicting instrument response functions(Camredon et al., 2007). Bauer and Grimmer (2016) recently reviewed the current performance of quantum chemistry methodologies in predicting EI mass spectrometry for small to medium sized molecules from first principles. Whilst that study documents improving general applicability, they are not immediately suitable for predicting AMS mass spectra because the thermal desorption promotes further fragmentation and, in some cases, pyrolysis(Canagaratna et al., 2015). While the standard AMS analysis takes these processes into account through empirical calibrations, the exact physical processes taking place within the vaporiser system are still the subject of considerable debate (Murphy, 2016;Drewnick et al., 2015;Robinson et al., 2016), so the bottom-up modelling of this is not possible with the current state of knowledge.

Distinct from all previous approaches, the approach presented here relies on supervised learning methods to automatically optimise the relationship between spectral characteristics and molecular features from the instrument in question. Therefore, any internal mechanisms or instrument features impacting on fragmentation are implicitly accounted for in the fitted model.

In section 2 the methodology behind constructing a predictive model is presented, whereas section 3 focuses on results regarding the accuracy of a model with respect to comparisons with spectra for individual components. In addition we present results from simulating the mass spectra of α-pinene aerosol using the GECKO-A model before we discuss future data requirements in section 4.

2 Methodology

Figure 1 displays the workflow used in building the predictive model. First, a model is trained to predict the occurrence of specific m/z channels as a function of molecular composition before a model for each m/z channel is trained to predict peak height within that channel. It is worthwhile detailing the molecular information used to train each model. Each molecule has varying levels of structural features, which can be written in terms of a 'fingerprint'. This fingerprint is a numerical identification of a given structure that can equally be thought of as stoichiometric information for distinct features. For example, for a collection of 10 compounds we would construct a matrix of stoichiometric information where each row represents a specific molecule and each column the stoichiometry of a given feature. We now refer to each column as a 'key', which might be a specific functional group or feature associated with that molecule. We retain the use of the word 'key' since it can provide more generic information than a functional group. To re-iterate, the entire row we refer to as the molecular fingerprint. For example, identifying the occurrence of carboxylic acid groups is a key within the AIOMFAC fingerprint (Zuend et al., 2011). We then take this information and use it to train a model to predict both the occurrence of a specific m/z channel and then peak heights.

To re-iterate, in constructing a model that can predict AMS mass spectra, a library of compounds with measured spectra are used to train a series of regression techniques. This collection of molecules, represented as SMILES strings, is parsed to produce a matrix where each column represents the stoichiometry of a particular key, or feature. This entire matrix is used to fit a predict model for each m/z channel.

The underlying physical principles of EI (F. W. McLafferty, 1994) adjusted to the AMS (Gasteiger et al., 1992), do not exist in algorithmic form, so there is currently no a priori basis for choosing the most appropriate fingerprint for this work. Therefore a collection of common fingerprints, and their combination, is tested in this study and their performance critically evaluated. This is an important sensitivity since one might expect a collection of keys that relate to EI fragmentation principles to offer a more robust basis for fitting any method used here. We discuss this further in section 4.

Fingerprints used in this study include those employed in activity coefficient and vapour pressure predictive techniques provided by the UManSysProp package (Topping et al., 2016;Zuend et al., 2011;Nannoolal et al., 2008), alongside more general fingerprints including the MACCS keys and FP4 keys (Putta et al., 2003). It is difficult to find information on provenance behind these latter generic fingerprints (Putta et al., 2003), other that they are designed to cover a set of molecular features that would be used across a broad range of applications. The MACCS fingerprint provides up-to 162 unique keys of any given molecule, the FP4 fingerprint featuring up to 320. The current implementation of the MACCS keys from the Pybel package (O'Boyle et al., 2011) is used whereas the FP4 keys are extracted from the RDKit open source informatics package (http://www.rdkit.org/docs/index.html). Each key is represented in the UManSysProp package (Topping et al., 2016) using SMARTS notation, and each molecule using the SMILES format. The matrix of keys used to fit each method is constructed by systematically parsing each molecule. Figure 2 demonstrates the use of the MACCS SMARTS to populate a matrix of keys. There are some common features between each fingerprint library, but also a range of differences. For example, all libraries identify the presence of the CH2 group, but then differ in the optional connecting groups. The FP4 keys cycle through systematic groupings, such as: primary carbon, secondary carbon, tertiary carbon…primary alcohol, secondary alcohol, tertiary alcohol etc. Similar groups are detected using the activity coefficient and vapour pressure keys. The full collection of SMARTS keys can be found in the source code and we discuss suggestions for future work on refining fingerprints in section 4. Please refer to section 5 on code availability.

With regards to the supervised methods used, an ensemble tree is trained to predict the occurrence of specific m/z channels as a function of any given fingerprint. To predict peak height per m/z channel, we evaluate a number of supervised methods available in the SciKit-learn package: Generalised Linear methods, Support Vector Machines [with 3 separate kernels], Stochastic Gradient Descent, Bayesian Ridge, Ordinary Least Squares, Decision Trees and Ensemble methods (Pedregosa et al., 2011). There are a number of other methods available yet, as we will discuss in section 4, the results from this study demonstrate a potential whilst further data is needed to confirm general applicability, including the use of other methods. For

a brief overview of each method, we refer the reader to Ruske (2016) and references therein. Before training each method, the matrix of identified keys were standardized between zero and one using the MinMaxScaler pre-processing feature within the Scikit learn package. In addition, the use of variable selection is designed to use only those features deemed important to construct fingerprint-peak height relationships to try and mitigate any under or over fitting. The sensitivity to these procedures are discussed in section 3.2. To compare modelled and measured mass spectra, the cosine angle from a dot product of the two are used, focusing on specific m/z channels that are typically found as features within atmospheric and smog chamber mass spectra (Ulbrich et al., 2009): 15,18,28,29,39,41,43,44,50,51,53,55,57,60,73,77,91.

The ability of each method to replicate the entire database is first evaluated. Whilst training on a subset and comparing with the entire database will test wider applicability, this initial comparison quantifies the appropriateness of the different fingerprints in building an accurate model.

3. Results

3.1 Sensitivity to choice of molecular fingerprint

Figure 3 visually compares the number of keys extracted from the 100 compounds in the AMS library according to choice of fingerprint. Data is presented according to the use of AIOMFAC [bottom left], MACCS [top left], Nanoolal [bottom right] and FP4 [top right] keys. Using the AIOMFAC fingerprint leads to, at most, 17 keys identified from the AMS library. The Nanoolal fingerprint leads to a larger set of keys (19), with the MACCS fingerprint providing the most (74) and the FP4 keys the second highest (30). The use of more or less information in the fitting procedure should not be assumed to automatically lead to a more accurate predictive model. Ideally there should be a balance between the number of features identified and how those features relate to the mechanisms of fragmentation on the molecule within the instrument in question. As we have already noted, comparing the information provided by each fingerprint with a working knowledge of the mechanics of EI fragmentation might help understand why a given fingerprint is more suitable. However we first and foremost wish to demonstrate the efficacy of using pre-defined fingerprints as they are available in the literature or within existing open-source software packages. The exact physical processes taking place within instrument are still the subject of considerable debate.

Table 1 presents the median cosine angle of modelled spectra fit to the entire AMS database derived from the different supervised methods and different fingerprints, either isolated or combined into one, to 2 decimal places. The left hand sided box-plots in figure 4a-d display the entire cosine angle spread for each method for the isolated MACCS (4a), FP4 (4b), AIOMFAC (4c) and Nanoolal fingerprints (4d). When fitting to the entire library of AMS spectra, initial results suggest that the tree-based methods [‘Tree’,’Forest’] perform better than others, with the MACCS keys leading to improved model performance over other fingerprints. However, the difference between using either the MACCS or Nanoolal keys, for

example, is not significant for any given supervised method as noted in Table 1. Rather than demonstrating 100% accuracy, the values of 1.00 must be taken with caution as we demonstrate in proceeding analyses. Whichever fingerprint is used, the ranking of performance between supervised methods remains similar, with the tree-based methods, Ordinary Least Squares and Bayesian Ridge outperforming Stochastic Gradient Descent and all Support Vector Machine kernels. Along with higher median values, the spread of cosine angles from the tree based methods and Ordinary Least Squares is much lower than all other methods. Whilst the use of MACCS and FP4 provide, in theory, more information, there is some similarity in structural information provided in all keys, as already discussed. For example, each fingerprint identifies key functional groups such as alkanes, alcohol, ketones etc, whilst the FP4 and MACCS keys in particular include more positional detail including relative positions of groups. At least for the 100 compounds in the AMS library, that additional information leads to a slight increase in cosine angle agreement of around 0.02 between methods, if we use only results from table 1 and figure 4. A key objective of this study, noted above, is to demonstrate the use of pre-defined fingerprints in constructing a predictive model. However, it is useful to also demonstrate the efficacy of combining the information from each fingerprint into one, without relating variable performance according to physical processes taking place within the instrument. The performance of combining all fingerprints into one, represented in table 1 under the column heading 'combined', illustrates a similar trend in performance between methods.

We discuss the significance of values displayed in table 1 after performance is re-evaluated following a more general approach of training to a subset of compounds, and the use of variable selection, in the next section.

3.2 Training to a subset, variable selection and dimensionality reduction

Table 2 presents the median cosine angle between modelled and predicted mass spectra, as a function of fingerprint, either isolated or combined into one, and regression technique, when training to a subset of the entire database and use of variable selection. To minimise over fitting any model to specific features, the process of variable selection allows us to refit the model to those keys deemed most important. The combination of both strategies might be considered the most suitable test of the methodology presented, with the full spread of statistics presented in the right hand column of figures 4a-d. It should be noted that randomly selecting the subset used for training leads to a significant decrease in model performance. This is due to missing keys within the training subset that are deemed important in predicting spectra for those compounds outside of the subset. A different approach is to select the subset according to maximising the number of keys across each molecule in the training subset, and is used in our proceeding analysis.

In some cases, such as with the Ordinary Least Squares and Forest methods, the data provided in Table 2 suggests that using both strategies leads to a lower median cosine angle, thus slightly reduced model performance when using isolated fingerprints. However, in practice, the statistics presented in Table 1 should not be considered a true test of the methodology,

but rather a precursor demonstration of the sensitivity to choice of fingerprint, and perhaps any variability in instrument response across the AMS library. On this, the use of the 'combined' fingerprint demonstrates the ability to retain information from those keys that improve overall performance.

Given their wide use across many disciplines, it is difficult to quantify the reasons behind the poor performance of the Support Vector Machines relative to other methods. To assess whether dimensional reduction procedures would improve accuracy, table 3 presents the median and overall spread of cosine angles when using Principal Component Analysis (PCA) on the 'combined' fingerprints. The number of principal components between 20, 10, 8 and 4. Generally, reducing the number of keys from up to 278 to 20 components, leads to an improvement of around 0.01-0.02 in all methods apart from Ordinary Least Squares and Support Vector Machines with both the polynomial and linear kernels. Results demonstrate clear sensitivity to the number of components when combined with the RBF Support Vector Machine kernel, performance varying from 0.84 to 0.67 on reducing the number of components from 20 to 4.

On the significance of the value of cosine angle, Figures 5 and 6 display predicted spectra for compounds not included in a training set, along with the cosine angle between modelled and measured spectra. From this point on we use isolated fingerprints to demonstrate the efficacy of our approach. For Oxalic acid, in Figure 5, the difference in performance between the FP4 and MACCS fingerprint [cosine of 0.83 and 0.77] is apparent through certain features, including the relative proportion of peak heights for the 3 dominate channels, and the ratio of f44 to f43. In Figure 6, a similar pattern is found for Leucine, including a marked difference in whether the model predicted non-zero entries across f41 – f44. Whilst a small subset, these results suggest use of the cosine angle alone is not sufficient to validate model performance, which is confirmed in section 3.3 when applied to the α-pinene system. Based on these comparisons, a tentative suggestion of using a cosine angle of 0.8 might go some way to clarifying the performance statistics provided in Tables 1 and 2 and Figure 4. Indeed, results demonstrate that, whilst statistics in Table 2 and Figure 4 suggest similar performance for both MACCS and FP4 keys, this performance is composition dependent. This reflects sensitivity to information used in the training process and how similarity between performances should be taken with caution in prescribing which method to take forward. This is better highlighted in the proceeding section with regards to a model SOA system.

Results at least suggest the tree based methods are at least the most stable given the higher range of cosine angles presented in Figures 4a-d and the decision tree method will be used in all proceeding analysis.

3.3 Example application to a model aerosol system.

In this section we apply the trained methods to a model SOA system, using output from the GECKO-A model used by (Valorso et al., 2011) to study SOA formation from α-pinene in a simulated chamber experiment. The purpose of this exercise is to explore sensitivity of predicted mass spectra to combined speciated output from a fixed model configuration

through varying fingerprints to support the comparisons made in the previous section. It is not designed as a thorough quantitative analysis of spectra comparisons, but rather to demonstrate the ability to extract specific features and highlight sensitivities to choice of model configuration. A recent study of McVay et al. (2016) presented results demonstrating sensitivity of aerosol mass and composition to processes included in a box-model model, including the addition of autoxidation mechanisms. They proposed that autoxidation might resolve some or all of measurement–model discrepancy from chamber simulations, but that this hypothesis could not be confirmed until more explicit mechanisms are established for α-pinene autoxidation(McVay et al., 2016). One might imagine an ideal sensitivity study would be to use speciated output from these updated models and add additional constraint to prescribing model performance through a comparison between measured and predicted mass spectra. Indeed, that is a rationale behind the study presented here. However, as proceeding results will demonstrate, with the existing training data and lack of validation on simple mixtures, there is potential for false positives in the predicted spectra to confuse a diagnosis of accurate model configurations. Specifically, the composition space derived from a series of box-model configurations would need to be mapped onto the existing space covered by the AMS spectral library. Combined with additional measurements of mixed systems of known composition, we could then prescribe a more robust set of regression model configurations through which a more detailed sensitivity study could take place.

Nonetheless, to illustrate sensitivity to choice of fingerprints in a complex system, Figure 7 displays the predicted mass spectra for the GECKO-A model results of Valorso et al. (2011) combined with the experimental data taken from a chamber-based α-pinene SOA formation experiment reported by Alfarra et al. (2013) (high VOC:NO_x ratio). Without further refinement of model and measurement conditions, these results exhibit large errors in the predicted mass spectra when using MACCS keys, despite the brief analysis presented in section 3.2. This demonstrates that over fitting to distinct features in the training set and difference between this composition space and that provided by the box-model output are leading to features that are missed in the final spectra. This is further supported by the abundance of features extracted from the training set displayed in figure 3.

To expand on this performance, Figure 8 displays the predicted mass spectra f44 peak height versus O:C ratio from the GECKO-A model results of Valorso et al (2011) in a manner similar to Aiken et al. (2008). There are 9 points on each curve, representing points in time during the GECKO-A simulation, with the model predicting a monotonic increase in O:C over time. It is worth noting the values are low compared to typical atmospheric LV-OOA (Aiken et al., 2008;Kroll et al., 2011). Overall, use of the FP4 and Nanoolal keys give absolute f44s that compare well with published calibrations relative to O:C, specifically Aiken et al. (2008) and the updated calibration presented by Canagaratna et al. (2015). The direction of the trend in f44 versus O:C is reversed when using the Nanoolal keys, with f44 decreasing with O:C, which runs contrary to expectations. However, it should be noted that the values are within the spread of values used to generate the Aiken et al.

(2008) and Canagaratna et al. (2015) calibrations, as these performed regressions over much bigger ranges of O:C than obtained in this simulation, so the prediction based on Nanoolal keys could still be plausible.

Figure 9 displays the predicted f44 to f43 peak heights from the model system using the commonly used 'triangle plot' (Morgan et al., 2010;Ng et al., 2011), compared with the experimental data taken from the chamber experiments of Alfarra et al. (2013) and also Chhabra et al. (2011), who studied the formation of α-pinene oxidation in response to different oxidants. Note the trajectories in this space are not monotonic for either the experimental or simulated data, which indicate the complexities in interpreting spectra based on these metrics. Results suggest that f43 values when using the FP4 and Nanoolal keys are plausible when compared to published studies. The f44 peak height is systematically low for all fingerprints, as also shown in figure 5-7. However, rather than a deficiency in the mass spectral prediction methods, this is likely due to a deficiency in the Valorso et al. (2011) model treatment. It has recently been shown how important mechanisms such as autooxidation are to the α-pinene SOA system (Ehn et al., 2014), which are capable of rapidly adding oxygenated functional groups to the molecules that are responsible for both the suppression of vapour pressures necessary for SOA formation and also the increase in the f44 metric (Canagaratna et al., 2015). More recent versions of GECKO-A have included such mechanisms (McVay et al., 2016), however a systematic comparison of the predicted spectra based on these inclusions is beyond the scope of this proof-of-concept paper and will be presented in a future publication.

4. Discussion and future work

The preceding analysis demonstrates the potential for the methodology presented to lead to interesting investigations on model versus measured mass spectra. However, there are a number of remaining improvements that need to be made. It is inevitable that not all of the chemical species predicted by the models will be covered by previous laboratory work. If a class of species predicted by any chemical mechanism is identified as not covered by existing SMARTS-based fragmentation rules, it could be characterised in the laboratory using the same facilities and methodologies employed for previous characterisation work (Canagaratna et al., (2015) and references therein).

On the sensitivity to choice of fingerprint, our results demonstrate compound specific trends that lead to performance variability when applied to a complex SOA system that is not apparent when analysing median cosine angle statistics. Combining available fingerprints into one can slightly improve performance in some cases, but as the comparison of isolated MACCS versus FP4 performance illustrates, there is potential danger in over fitting to distinct features in the training set that is not provided by the box-model output. To re-iterate, one might expect a collection of keys that relate to EI fragmentation principles to offer a more robust basis for fitting any method used here. However, that requires further work with additional laboratory data to validate the efficacy of any new bespoke fingerprint.

The methods here have a number of uses, although it must be re-iterated that the predicted mass spectra are not definitive. The performance of this method will be improved by the addition of further training data. Following the development of group contribution methods, this could include studies on compounds within a specific series and mixtures of those compounds. As outlined in the introduction, the ability of this model to predict AMS spectra will be useful in the development and validation of explicit SOA mechanisms in the laboratory, meaning that the models can be challenged by the entire mass spectrum and not just the mass and O:C ratio. This method can also be used at the experiment design stage, allowing predictions of whether an AMS will be able to discern expected changes in composition associated with a process and thus whether it will be useful to test particular hypotheses.

The method could also be used to simulate atmospheric aerosol, probably if the chemical model is used in a Lagrangian configuration. In addition to the insights gained in atmospheric processes, this could be used to critically test the data model used in positive matrix factorisation (PMF) (Ulbrich et al., 2009). Because of the condition that PMF factors have fixed profiles, the reduction of the complexity associated with atmospheric SOA to (typically) two factors results in an increase in 'rotational ambiguity' associated with the factorisation. A two-component factorisation of SOA is often interpreted as representing the 'low volatility' and 'semivolalite' components of the SOA (Jimenez et al., 2009), although this has shown not to be applicable to all environments, where other sources of variability contribute to the split in the factors (Young et al., 2015). If the mass spectral response to atmospheric SOA could be more explicitly simulated using this technique, a synthetic AMS dataset could be used as the subject of PMF analysis in a manner similar to Ulbrich et al. (2009). This in turn could be used to investigate the contributions of the factorisation on a more explicit level and investigate the effects this has on rotational ambiguity and the validity of solutions.

5. Code availability

A publicly available copy of the code used to derive performance statistics of the chosen regression methods can be found at : https://github.com/loftytopping/STRAPS covered by a GPL v3.0 license. This includes a copy of the AMS spectral files that now also include appropriate SMILEs strings. The code separates the four fingerprint libraries used in this study. We also provide an associated DOI for the exact model version given in this paper as provided by the Zenodo service: https://zenodo.org/record/213068#.WFlryyiPD3s

Please note that an extension to the SMARTS libraries included in UmanSysProp was carried out in this project. To review the features extracted for each fingerprint, please refer to the files 'FP4.smarts', 'MACCS.smarts', 'nannoolal_primary.smarts' and 'aiomfac_unifac.smarts' included in the directory *UManSysProp_public/umansysprop/data/*.

Author contributions:

David Topping conceived the methodology presented and performed the subsequent model development and analysis. James Allan and Rami Alfarra offered expert guided constraints on evaluating the model results, including selecting the best comparison metrics to use. Bernard Aumont supplied the results from the Valorso et al (2011) study. All authors contributed to the writing of the manuscript.

Acknowledgements: David Topping, James Allan and Rami Alfarra received funding from the National Centre for Atmospheric Science [NCAS]. This work was built on informatics developed under NERC grant NE/H002588/1.

References:

Aiken, A. C., DeCarlo, P. F., and Jimenez, J. L.: Elemental analysis of organic species with electron ionization high-resolution mass spectrometry, Anal Chem, 79, 8350-8358, 10.1021/ac071150w, 2007.

Aiken, A. C., Decarlo, P. F., Kroll, J. H., Worsnop, D. R., Huffman, J. A., Docherty, K. S., Ulbrich, I. M., Mohr, C., Kimmel, J. R., Sueper, D., Sun, Y., Zhang, Q., Trimborn, A., Northway, M., Ziemann, P. J., Canagaratna, M. R., Onasch, T. B., Alfarra, M. R., Prevot, A. S. H., Dommen, J., Duplissy, J., Metzger, A., Baltensperger, U., and Jimenez, J. L.: O/C and OM/OC ratios of primary, secondary, and ambient organic aerosols with high-resolution time-of-flight aerosol mass spectrometry, Environ Sci Technol, 42, 4478-4485, 10.1021/es703009q, 2008.

Alfarra, M. R., Good, N., Wyche, K. P., Hamilton, J. E., Monks, P. S., Lewis, A. C., and McFiggans, G.: Water uptake is independent of the inferred composition of secondary aerosols derived from multiple biogenic VOCs, Atmos Chem Phys, 13, 11769-11789, 10.5194/acp-13-11769-2013, 2013.

Aumont, B., Szopa, S., and Madronich, S.: Modelling the evolution of organic carbon during its gas-phase tropospheric oxidation: development of an explicit model based on a self generating approach, Atmos Chem Phys, 5, 2497-2517, 2005.

Aumont, B., Valorso, R., Mouchel-Vallon, C., Camredon, M., Lee-Taylor, J., and Madronich, S.: Modeling SOA formation from the oxidation of intermediate volatility n-alkanes, Atmos Chem Phys, 12, 7577-7589, 10.5194/acp-12-7577-2012, 2012.

Barley, M. H., and McFiggans, G.: The critical assessment of vapour pressure estimation methods for use in modelling the formation of atmospheric organic aerosol, Atmos Chem Phys, 10, 749-767, 10.5194/acp-10-749-2010, 2010.

Bauer, C. A., and Grimme, S.: How to Compute Electron Ionization Mass Spectra from First Principles, J Phys Chem A, 120, 3755-3766, 10.1021/acs.jpca.6b02907, 2016.

Bergstrom, R., van der Gon, H. A. C. D., Prevot, A. S. H., Yttri, K. E., and Simpson, D.: Modelling of organic aerosols over Europe (2002-2007) using a volatility basis set (VBS) framework: application of different assumptions regarding the formation of secondary organic aerosol, Atmos Chem Phys, 12, 8499-8527, 10.5194/acp-12-8499-2012, 2012.

Bilde, M., Barsanti, K., Booth, M., Cappa, C. D., Donahue, N. M., Emanuelsson, E. U., McFiggans, G., Krieger, U. K., Marcolli, C., Tropping, D., Ziemann, P., Barley, M., Clegg, S., Dennis-Smither, B., Hallquist, M., Hallquist, A. M., Khlystov, A., Kulmala, M., Mogensen, D., Percival, C. J., Pope, F., Reid, J. P., da Silva, M. A. V. R., Rosenoern, T., Salo, K., Soonsin, V. P., Yli-Juuti, T., Prisle, N. L., Pagels, J., Rarey, J., Zardini, A. A., and Riipinen, I.: Saturation Vapor Pressures and Transition Enthalpies of Low-Volatility Organic Molecules of Atmospheric Relevance: From Dicarboxylic Acids to Complex Mixtures, Chem Rev, 115, 4115-4156, 10.1021/cr5005502, 2015.

Camredon, M., Aumont, B., Lee-Taylor, J., and Madronich, S.: The SOA/VOC/NOx system: an explicit model of secondary organic aerosol formation, Atmos Chem Phys, 7, 5599-5610, 2007.

Canagaratna, M. R., Jimenez, J. L., Kroll, J. H., Chen, Q., Kessler, S. H., Massoli, P., Hildebrandt Ruiz, L., Fortner, E., Williams, L. R., Wilson, K. R., Surratt, J. D., Donahue, N. M., Jayne, J. T., and Worsnop, D. R.: Elemental ratio measurements of organic compounds using aerosol mass spectrometry: characterization, improved calibration, and implications, Atmos Chem Phys, 15, 253-272, 10.5194/acp-15-253-2015, 2015.

Chhabra, P. S., Ng, N. L., Canagaratna, M. R., Corrigan, A. L., Russell, L. M., Worsnop, D. R., Flagan, R. C., and Seinfeld, J. H.: Elemental composition and oxidation of chamber organic aerosol, Atmos Chem Phys, 11, 8827-8845, 10.5194/acp-11-8827-2011, 2011.

Drewnick, F., Diesch, J. M., Faber, P., and Borrmann, S.: Aerosol mass spectrometry: particle-vaporizer interactions and their consequences for the measurements, Atmos Meas Tech, 8, 3811-3830, 10.5194/amt-8-3811-2015, 2015.

Ehn, M., Thornton, J. A., Kleist, E., Sipila, M., Junninen, H., Pullinen, I., Springer, M., Rubach, F., Tillmann, R., Lee, B., Lopez-Hilfiker, F., Andres, S., Acir, I. H., Rissanen, M., Jokinen, T., Schobesberger, S., Kangasluoma, J., Kontkanen, J., Nieminen, T., Kurten, T., Nielsen, L. B., Jorgensen, S., Kjaergaard, H. G., Canagaratna, M., Dal Maso, M., Berndt, T., Petaja, T., Wahner, A., Kerminen, V. M., Kulmala, M., Worsnop, D. R., Wildt, J., and Mentel, T. F.: A large source of low-volatility secondary organic aerosol, Nature, 506, 476-+, 10.1038/nature13032, 2014.

F. W. McLafferty, F. T.: Interpretation of mass spectra, edited by: Vetter, W., University Science Books, Mill Valley, California, 1994.

Gasteiger, J., Hanebeck, W., and Schulz, K. P.: Prediction of Mass-Spectra from Structural Information, J Chem Inf Comp Sci, 32, 264-271, Doi 10.1021/Ci00008a001, 1992.

Hallquist, M., Wenger, J. C., Baltensperger, U., Rudich, Y., Simpson, D., Claeys, M., Dommen, J., Donahue, N. M., George, C., Goldstein, A. H., Hamilton, J. F., Herrmann, H., Hoffmann, T., Iinuma, Y., Jang, M., Jenkin, M. E., Jimenez, J. L., Kiendler-Scharr, A., Maenhaut, W., McFiggans, G., Mentel, T. F., Monod, A., Prevot, A. S. H., Seinfeld, J. H., Surratt, J. D., Szmigielski, R., and Wildt, J.: The formation, properties and impact of secondary organic aerosol: current and emerging issues, Atmos Chem Phys, 9, 5155-5236, 2009.

Hamilton, J. F., Baeza-Romero, M. T., Finessi, E., Rickard, A. R., Healy, R. M., Peppe, S., Adams, T. J., Daniels, M. J. S., Ball, S. M., Goodall, I. C. A., Monks, P. S., Borras, E., and Munoz, A.: Online and offline mass spectrometric study of the impact of oxidation and ageing on glyoxal chemistry and uptake onto ammonium sulfate aerosols, Faraday Discuss, 165, 447-472, 10.1039/c3fd00051f, 2013.

Jimenez, J. L., Canagaratna, M. R., Donahue, N. M., Prevot, A. S. H., Zhang, Q., Kroll, J. H., DeCarlo, P. F., Allan, J. D., Coe, H., Ng, N. L., Aiken, A. C., Docherty, K. S., Ulbrich, I. M., Grieshop, A. P., Robinson, A. L., Duplissy, J., Smith, J. D., Wilson, K. R., Lanz, V. A., Hueglin, C., Sun, Y. L., Tian, J., Laaksonen, A., Raatikainen, T., Rautiainen, J., Vaattovaara, P., Ehn, M., Kulmala, M., Tomlinson, J. M., Collins, D. R., Cubison, M. J., Dunlea, E. J., Huffman, J. A., Onasch, T. B., Alfarra, M. R., Williams, P. I., Bower, K., Kondo, Y., Schneider, J., Drewnick, F., Borrmann, S., Weimer, S., Demerjian, K., Salcedo, D., Cottrell, L., Griffin, R., Takami, A., Miyoshi, T., Hatakeyama, S., Shimono, A., Sun, J. Y., Zhang, Y. M., Dzepina, K., Kimmel, J. R., Sueper, D., Jayne, J. T., Herndon, S. C., Trimborn, A. M., Williams, L. R., Wood, E. C., Middlebrook, A. M., Kolb, C. E., Baltensperger, U., and Worsnop, D. R.: Evolution of Organic Aerosols in the Atmosphere, Science, 326, 1525-1529, 10.1126/science.1180353, 2009.

Kroll, J. H., Donahue, N. M., Jimenez, J. L., Kessler, S. H., Canagaratna, M. R., Wilson, K. R., Altieri, K. E., Mazzoleni, L. R., Wozniak, A. S., Bluhm, H., Mysak, E. R., Smith, J. D., Kolb, C. E., and Worsnop, D. R.: Carbon oxidation state as a metric for describing the chemistry of atmospheric organic aerosol, Nat Chem, 3, 133-139, 10.1038/NCHEM.948, 2011.

Laaksonen, A., Kulmala, M., O'Dowd, C. D., Joutsensaari, J., Vaattovaara, P., Mikkonen, S., Lehtinen, K. E. J., Sogacheva, L., Dal Maso, M., Aalto, P., Petaja, T., Sogachev, A., Yoon, Y. J., Lihavainen, H., Nilsson, D., Facchini, M. C., Cavalli, F., Fuzzi, S., Hoffmann, T., Arnold, F., Hanke, M., Sellegri, K., Umann, B., Junkermann, W., Coe, H., Allan, J. D., Alfarra, M. R., Worsnop, D. R., Riekkola, M. L., Hyotylainen, T., and Viisanen, Y.: The role of VOC oxidation products in continental new particle formation, Atmos Chem Phys, 8, 2657-2665, 2008.

McVay, R. C., Zhang, X., Aumont, B., Valorso, R., Camredon, M., La, Y. S., Wennberg, P. O., and Seinfeld, J. H.: SOA formation from the photooxidation of alpha-pinene: systematic exploration of the simulation of chamber data, Atmos Chem Phys, 16, 2785-2802, 10.5194/acp-16-2785-2016, 2016.

Morgan, W. T., Allan, J. D., Bower, K. N., Highwood, E. J., Liu, D., McMeeking, G. R., Northway, M. J., Williams, P. I., Krejci, R., and Coe, H.: Airborne measurements of the spatial distribution of aerosol chemical composition across Europe and evolution of the organic fraction, Atmos Chem Phys, 10, 4065-4083, 10.5194/acp-10-4065-2010, 2010.

Murphy, D. M.: The effects of molecular weight and thermal decomposition on the sensitivity of a thermal desorption aerosol mass spectrometer, Aerosol Sci Tech, 50, 118-125, 10.1080/02786826.2015.1136403, 2016.

Nannoolal, Y., Rarey, J., and Ramjugernath, D.: Estimation of pure component properties - Part 3. Estimation of the vapor pressure of non-electrolyte organic compounds via group contributions and group interactions, Fluid Phase Equilibr, 269, 117-133, 10.1016/j.fluid.2008.04.020, 2008.

Ng, N. L., Canagaratna, M. R., Jimenez, J. L., Chhabra, P. S., Seinfeld, J. H., and Worsnop, D. R.: Changes in organic aerosol composition with aging inferred from aerosol mass spectra, Atmos Chem Phys, 11, 6465-6474, 10.5194/acp-11-6465-2011, 2011.

Noziere, B., Kaberer, M., Claeys, M., Allan, J., D'Anna, B., Decesari, S., Finessi, E., Glasius, M., Grgic, I., Hamilton, J. F., Hoffmann, T., Iinuma, Y., Jaoui, M., Kahno, A., Kampf, C. J., Kourtchev, I., Maenhaut, W., Marsden, N., Saarikoski, S., Schnelle-Kreis, J., Surratt, J. D., Szidat, S., Szmigielski, R., and Wisthaler, A.: The Molecular Identification of Organic Compounds in the Atmosphere: State of the Art and Challenges, Chem Rev, 115, 3919-3983, 10.1021/cr5003485, 2015.

O'Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T., and Hutchison, G. R.: Open Babel: An open chemical toolbox, J Cheminformatics, 3, Artn 33
10.1186/1758-2946-3-33, 2011.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E.: Scikit-learn: Machine Learning in Python, J Mach Learn Res, 12, 2825-2830, 2011.

Putta, S., Eksterowicz, J., Lemmen, C., and Stanton, R.: A novel subshape molecular descriptor, J Chem Inf Comp Sci, 43, 1623-1635, 10.1021/ci0256384, 2003.

Robinson, A. L., Donahue, N. M., Shrivastava, M. K., Weitkamp, E. A., Sage, A. M., Grieshop, A. P., Lane, T. E., Pierce, J. R., and Pandis, S. N.: Rethinking organic aerosols: Semivolatile emissions and photochemical aging, Science, 315, 1259-1262, 10.1126/science.1133061, 2007.

Robinson, E. S., Donahue, N. M., Ahern, A. T., Ye, Q., and Lipsky, E.: Single-particle measurements of phase partitioning between primary and secondary organic aerosols, Faraday Discuss, 189, 31-49, 10.1039/c5fd00214a, 2016.

Ruske, S., Topping, D. O., Foot, V. E., Kaye, P. H., Stanley, W. R., Crawford, I., Morse, A. P., and Gallagher, M. W.: Evaluation of Machine Learning Algorithms for Classification of Primary Biological Aerosol using a new UV-LIF spectrometer, Atmos. Meas. Tech. Discuss, 10.5194/amt-2016-214, 2016.

Saunders, S. M., Jenkin, M. E., Derwent, R. G., and Pilling, M. J.: World Wide Web site of a Master Chemical Mechanism (MCM) for use in tropospheric chemistry models, Atmos Environ, 31, 1249-1249, Doi 10.1016/S1352-2310(97)85197-7, 1997.

Sipila, M., Sarnela, N., Jokinen, T., Henschel, H., Junninen, H., Kontkanen, J., Richters, S., Kangasluoma, J., Franchin, A., Perakyla, O., Rissanen, M. P., Ehn, M., Vehkamaki, H., Kurten, T., Berndt, T., Petaja, T., Worsnop, D., Ceburnis, D., Kerminen, V. M., Kulmala, M., and O'Dowd, C.: Molecular-scale evidence of aerosol particle formation via sequential addition of HIO3, Nature, 537, 532-534, 10.1038/nature19314, 2016.

Spracklen, D. V., Jimenez, J. L., Carslaw, K. S., Worsnop, D. R., Evans, M. J., Mann, G. W., Zhang, Q., Canagaratna, M. R., Allan, J., Coe, H., McFiggans, G., Rap, A., and Forster, P.: Aerosol mass spectrometer constraint on the global secondary organic aerosol budget, Atmos Chem Phys, 11, 12109-12136, 10.5194/acp-11-12109-2011, 2011.

Topping, D., Barley, M., Bane, M. K., Higham, N., Aumont, B., Dingle, N., and McFiggans, G.: UManSysProp v1.0: an online and open-source facility for molecular property prediction and atmospheric aerosol calculations, Geosci Model Dev, 9, 899-914, 10.5194/gmd-9-899-2016, 2016.

Ulbrich, I. M., Canagaratna, M. R., Zhang, Q., Worsnop, D. R., and Jimenez, J. L.: Interpretation of organic components from Positive Matrix Factorization of aerosol mass spectrometric data, Atmos Chem Phys, 9, 2891-2918, 10.5194/acp-9-2891-2009, 2009.

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 alpha-pinene photooxidation: sensitivity to vapour pressure estimation, Atmos Chem Phys, 11, 6895-6910, 10.5194/acp-11-6895-2011, 2011.

Volkamer, R., Jimenez, J. L., San Martini, F., Dzepina, K., Zhang, Q., Salcedo, D., Molina, L. T., Worsnop, D. R., and Molina, M. J.: Secondary organic aerosol formation from anthropogenic air pollution: Rapid and higher than expected, Geophys Res Lett, 33, Artn L17811
10.1029/2006gl026899, 2006.

Young, D. E., Allan, J. D., Williams, P. I., Green, D. C., Harrison, R. M., Yin, J., Flynn, M. J., Gallagher, M. W., and Coe, H.: Investigating a two-component model of solid fuel organic aerosol in London: processes, PM1 contributions, and seasonality, Atmos Chem Phys, 15, 2429-2443, 10.5194/acp-15-2429-2015, 2015.
Zuend, A., Marcolli, C., Booth, A. M., Lienhard, D. M., Soonsin, V., Krieger, U. K., Topping, D. O., McFiggans, G., Peter, T., and Seinfeld, J. H.: New and extended parameterization of the thermodynamic model AIOMFAC: calculation of activity coefficients for organic-inorganic mixtures containing carboxyl, hydroxyl, carbonyl, ether, ester, alkenyl, alkyl, and aromatic functional groups, Atmos Chem Phys, 11, 9155-9206, 10.5194/acp-11-9155-2011, 2011.

Method	MACCS	FP4	AIOMFAC	Nanoolal	Combined
SVM RBF	0.87	0.85	0.86	0.85	0.85
SVM Poly	0.84	0.83	0.82	0.81	0.83
SVM Lin	0.80	0.80	0.79	0.79	0.80
BRR	0.94	0.92	0.90	0.91	0.95
OLS	1.00	0.96	0.94	0.94	0.99
SGDR	0.88	0.82	0.80	0.80	0.89
Tree	1.00	1.00	1.00	1.00	1.00
Forest	1.00	1.00	1.00	1.00	1.00

Table1 - Median cosine angle between measured and predicted spectra when fitting to the entire dataset as a function of molecular fingerprint [Given above each column]. Please note, the term 'Combined' refers to a combination of all individual fingerprints into one. The method labels are as follows: SMV [Support vector Machine with 3 kernels (RBF, Poly[nomial] and Lin[near])], BRR: Bayesian Ridge, OLS: Ordinary Least Squares, SGDR:Stochastic Gradient Descent, Tree: Decision Tree and Forest: Random Forest.

Method	MACCS	FP4	AIOMFAC	Nanoolal	Combined
SVM RBF	0.85	0.82	0.80	0.81	0.85
SVM Poly	0.82	0.81	0.81	0.79	0.82
SVM Lin	0.78	0.79	0.78	0.78	0.80
BRR	0.93	0.91	0.88	0.88	0.94
OLS	0.95	0.93	0.90	0.90	0.98
SGDR	0.87	0.82	0.81	0.80	0.88
Tree	0.97	0.97	0.94	0.96	0.98
Forest	0.97	0.97	0.95	0.96	0.98

Table 2 - Median cosine angle between measured and predicted spectra, using 80% of the compounds in the training process, with variable selection, as a function of molecular fingerprint [Given above each column]. Please note, the term 'Combined' refers to a combination of all individual fingerprints into one. The method labels are as follows: SMV [Support vector Machine with 3 kernels (RBF, Poly[nomial] and Lin[near])], BRR: Bayesian Ridge, OLS: Ordinary Least Squares, SGDR:Stochastic Gradient Descent, Tree: Decision Tree and Forest: Random Forest.

Method	20	10	8	4
SVM RBF	0.84	0.84	0.85	0.67
SVM Poly	0.83	0.83	0.81	0.79
SVM Lin	0.80	0.80	0.80	0.80
BRR	0.93	0.90	0.89	0.87
OLS	0.94	0.89	0.89	0.87
SGDR	0.89	0.89	0.89	0.88
Tree	0.98	0.98	0.98	0.98
Forest	0.99	0.99	0.99	0.99

Table 3 - Median cosine angle between measured and predicted spectra, applying PCA analysis to the 'combined' fingerprints, as a function of the number of principal components used given above each column. The method labels are as follows: SMV [Support vector Machine with 3 kernels (RBF, Poly[nomial] and Lin[near])], BRR: Bayesian Ridge, OLS: Ordinary Least Squares, SGDR:Stochastic Gradient Descent, Tree: Decision Tree and Forest: Random Forest.

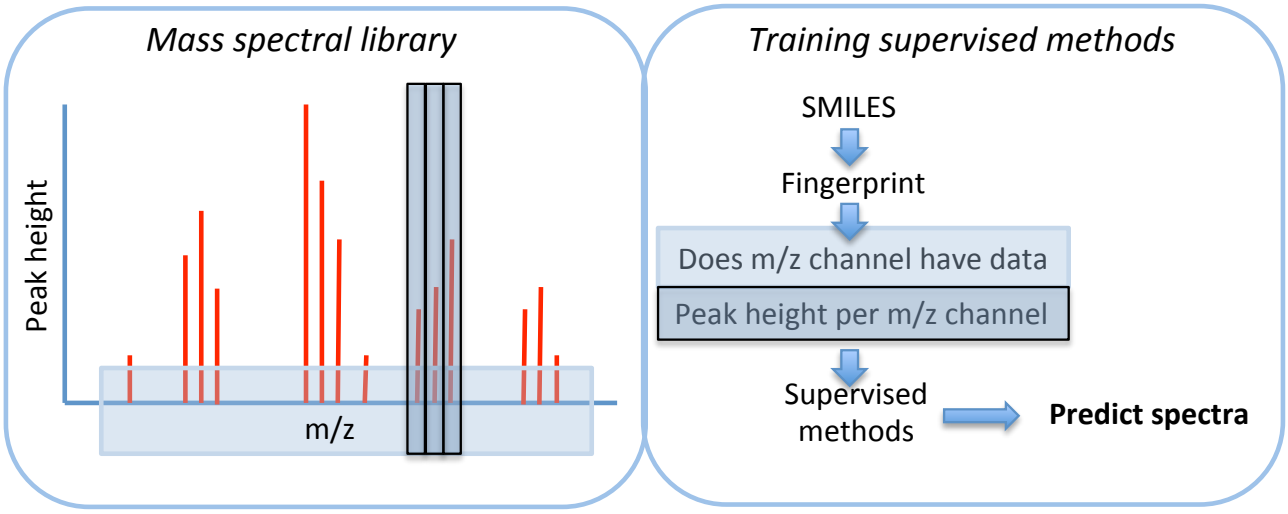

Figure 1 – Schematic of workflow used in the training process. For a normalised mass spectrum, the SMILEs string associated with each compound is combined with a given molecular fingerprint to train methods to predict the occurrence of a given m/z channel and then a peak height.

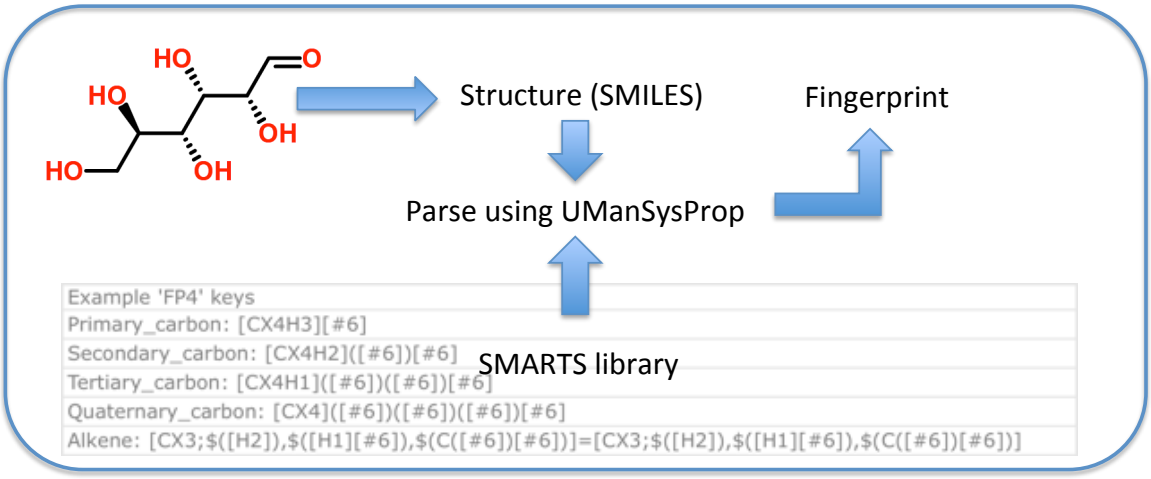

Figure 2. Basic schematic of interrogating a SMILES string with a SMARTS library to construct a molecular fingerprint.

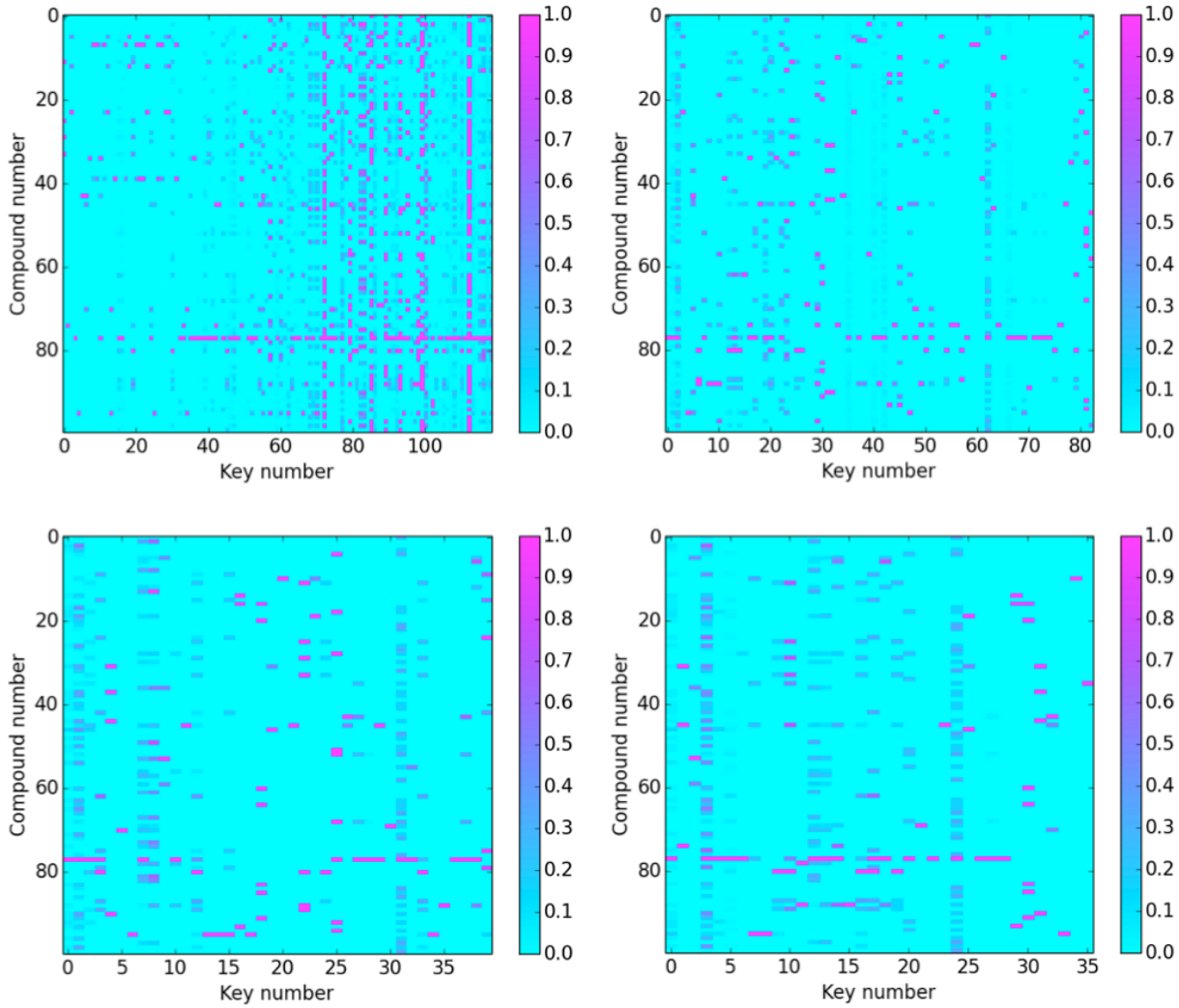

Figure 3 – Sparsity of keys extracted (x axes) from each compound (y axes) as a function of molecular fingerprint used (Top left: MACCS, Top right: FP4, Bottom left: AIOMFAC, Bottom right: Nanoolal). Keys are coloured according to normalised stoichiometry across all compounds.

5

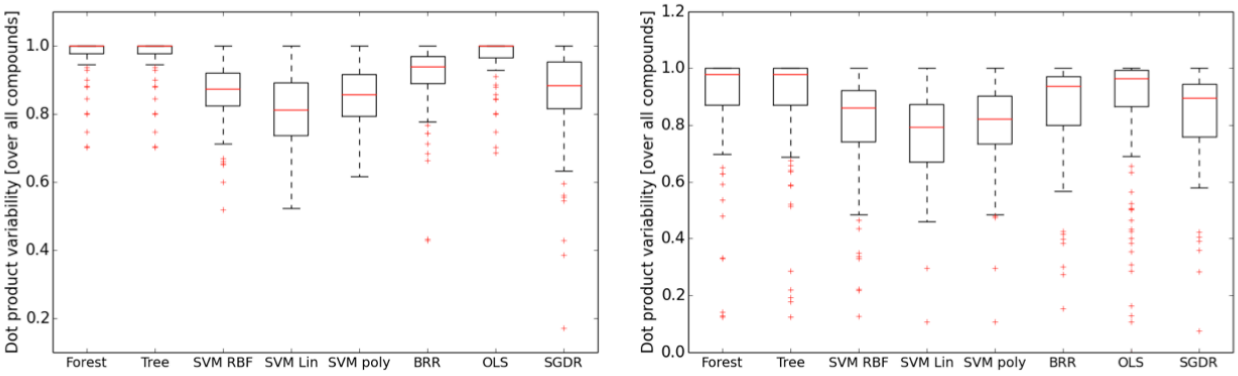

Figure 4a – Spread of cosine angle between experimental and predicted mass spectra [y axes] for all 100 compounds in the AMS library as a function of supervised method [x axes] using the MACCS fingerprint. left: using all compounds in the training process. right: using 80% of the compounds in the training process with variable selection. The method labels are as follows: SMV [Support vector Machine with 3 kernels (RBF, Poly[nomial] and Lin[near])], BRR: Bayesian Ridge, OLS: Ordinary Least Squares, SGDR:Stochastic Gradient Descent, Tree: Decision Tree and Forest: Random Forest.

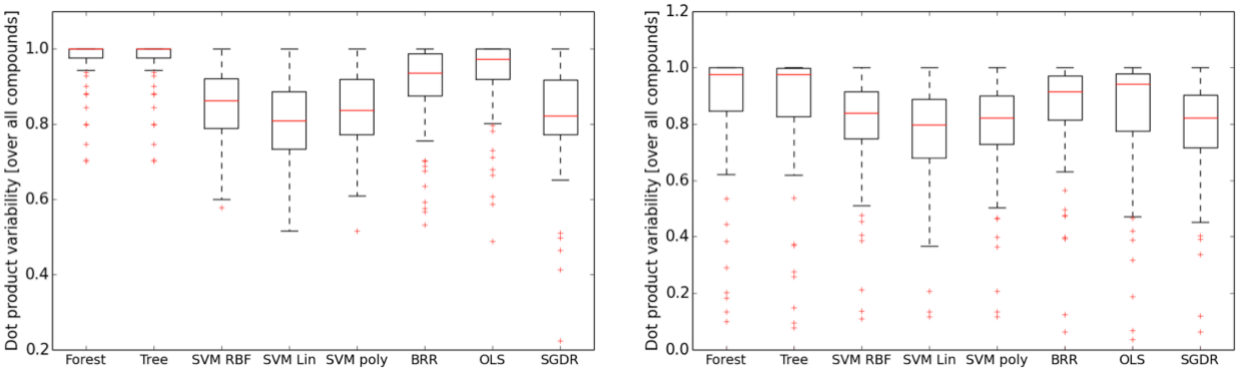

Figure 4b – Spread of cosine angle between experimental and predicted mass spectra [y axes] for all 100 compounds in the AMS library as a function of supervised method [x axes] using the FP4 fingerprint. left: using all compounds in the training process. right: using 80% of the compounds in the training process with variable selection. The method labels are as follows: SMV [Support vector Machine with 3 kernels (RBF, Poly[nomial] and Lin[near])], BRR: Bayesian Ridge, OLS: Ordinary Least Squares, SGDR:Stochastic Gradient Descent, Tree: Decision Tree and Forest: Random Forest.

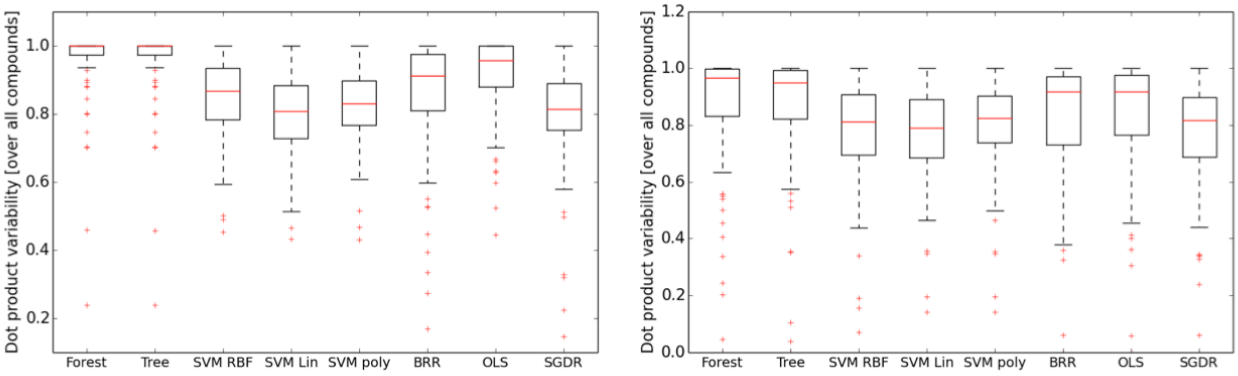

Figure 4c – Spread of cosine angle between experimental and predicted mass spectra [y axes] for all 100 compounds in the AMS library as a function of supervised method [x axes] using the AIOMFAC fingerprint. left: using all compounds in the training process. right: using 80% of the compounds in the training process with variable selection. The method labels are as follows: SMV [Support vector Machine with 3 kernels (RBF, Poly[nomial] and Lin[near])], BRR: Bayesian Ridge, OLS: Ordinary Least Squares, SGDR:Stochastic Gradient Descent, Tree: Decision Tree and Forest: Random Forest.

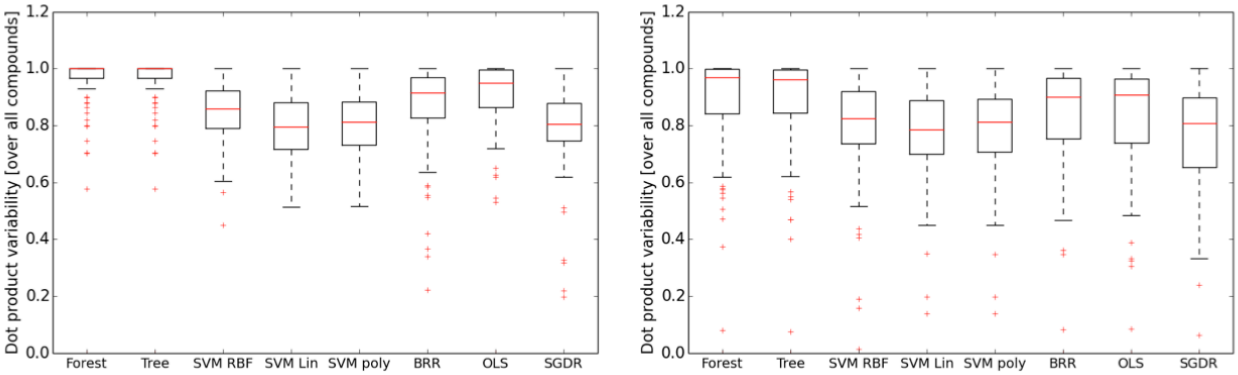

Figure 4d – Spread of cosine angle between experimental and predicted mass spectra [y axes] for all 100 compounds in the AMS library as a function of supervised method [x axes] using the Nanoolal fingerprint. left: using all compounds in the training process. right: using 80% of the compounds in the training process with variable selection. The method labels are as follows: SMV [Support vector Machine with 3 kernels (RBF, Poly[nomial] and Lin[near]), BRR: Bayesian Ridge, OLS: Ordinary Least Squares, SGDR:Stochastic Gradient Descent, Tree: Decision Tree and Forest: Random Forest.

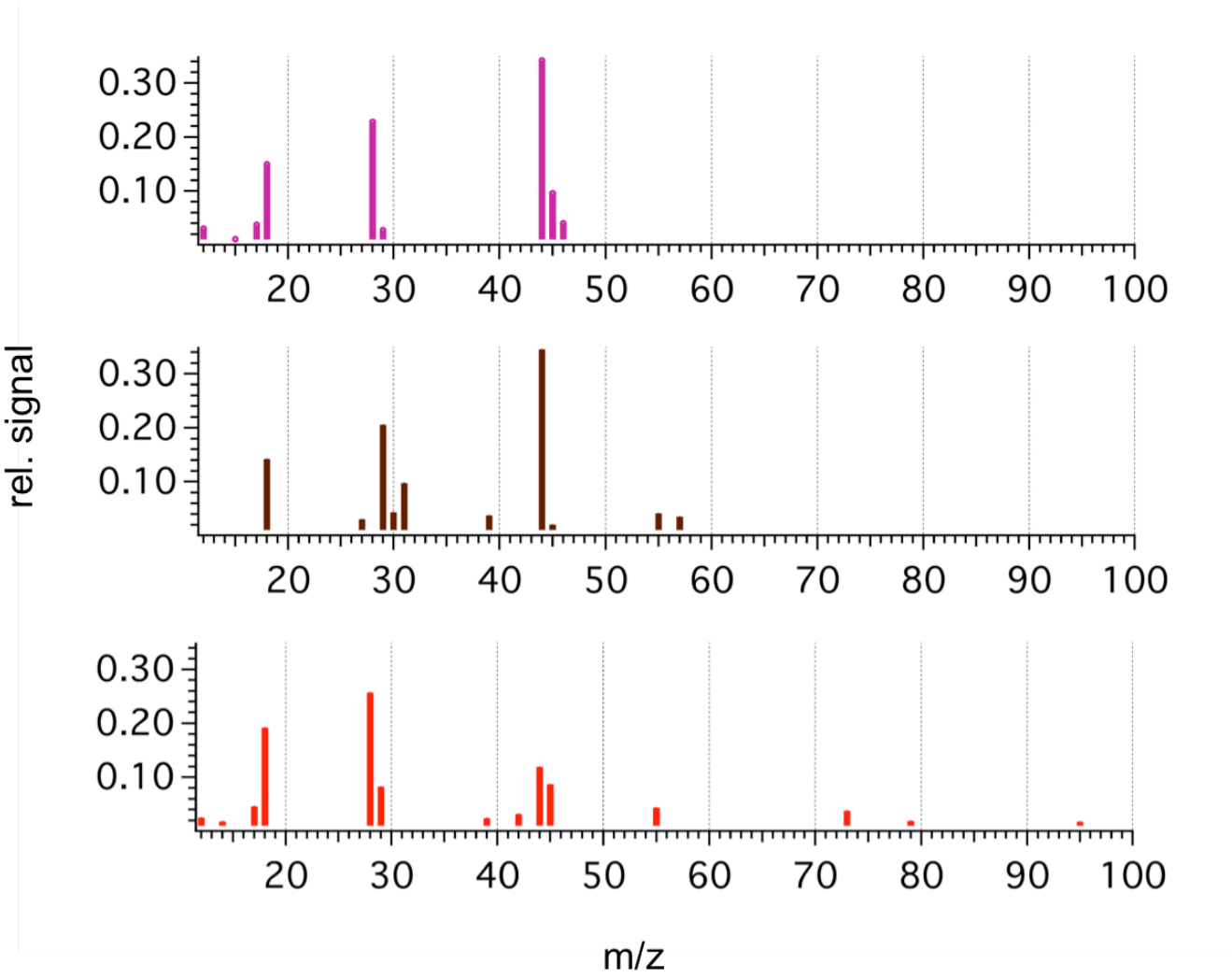

Figure 5 – Measured mass spectra for Oxalic acid [top] versus predicted mass spectra from an ensemble tree using the FP4 fingerprint [middle, cosine of 0.83] and the MACCS fingerprint [bottom, cosine of 0.77].

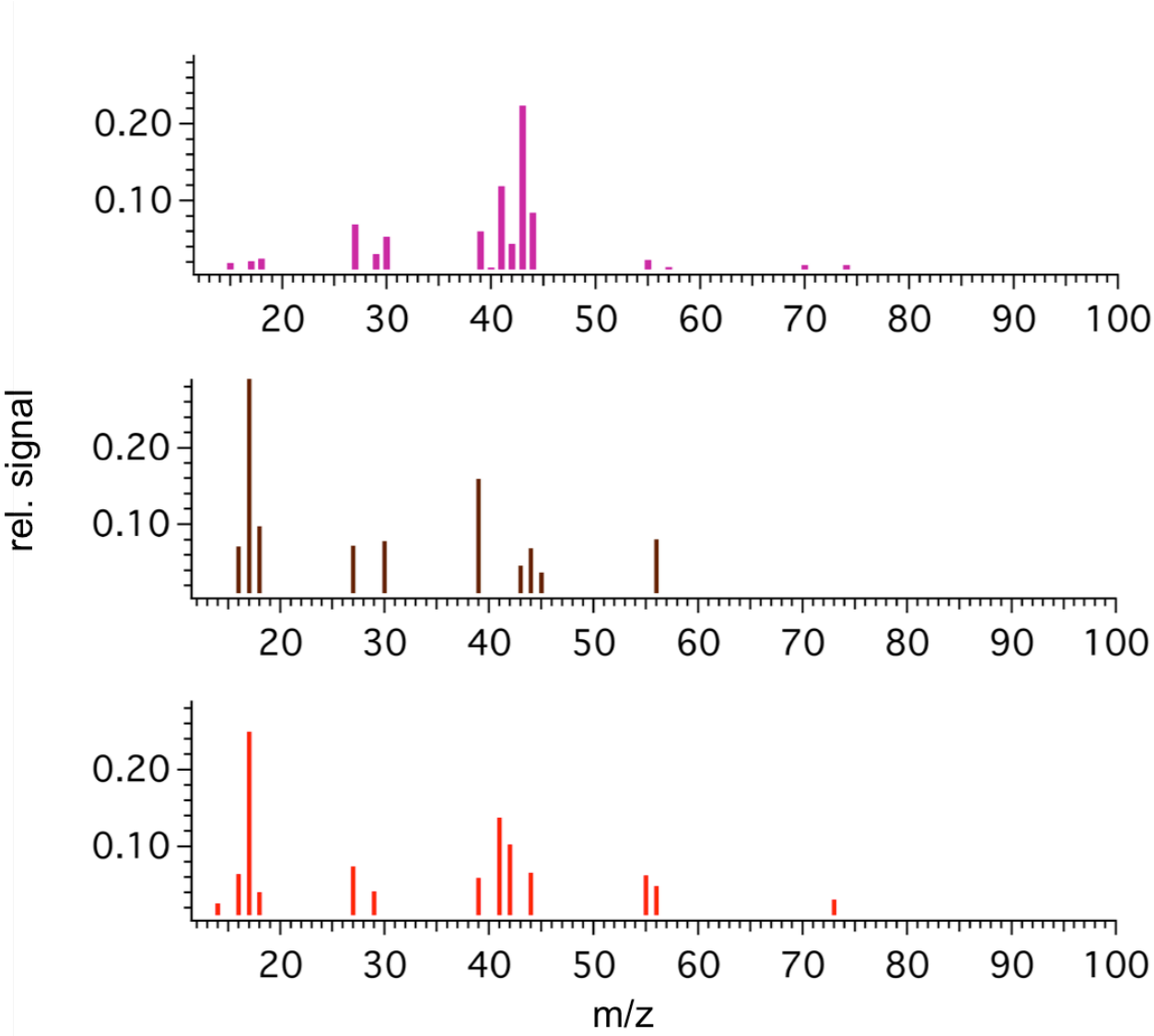

Figure 6 – Measured mass spectra for Leucine [top] versus predicted mass spectra from an ensemble tree using the FP4 fingerprint [middle, cosine of 0.70] and the MACCS fingerprint [bottom, cosine of 0.94].

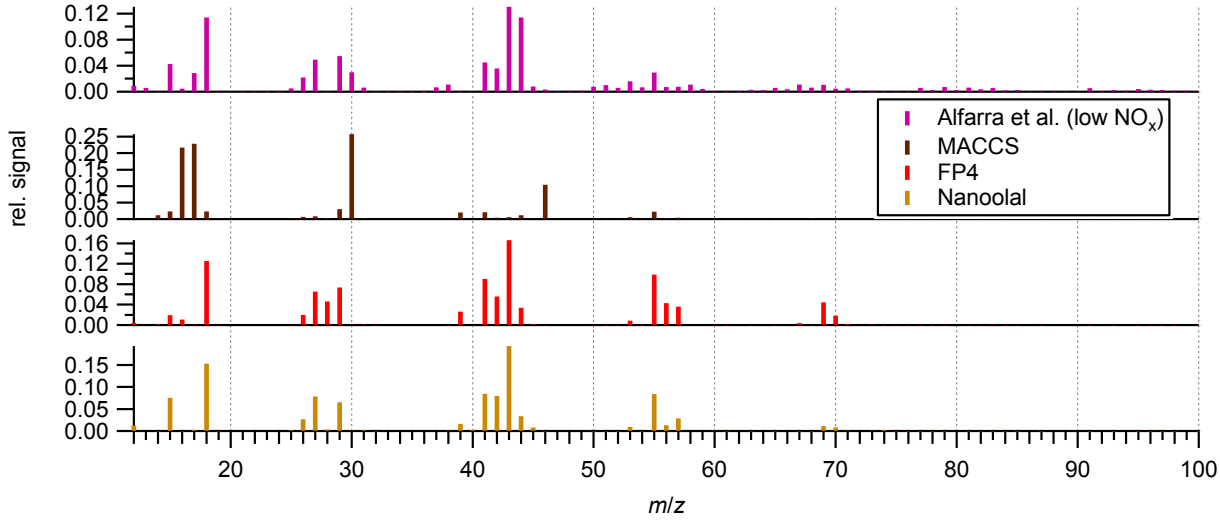

Figure 7 – Comparison of the predicted mass spectra of α-pinene SOA based on the GECKO-A simulation presented by Valorso et al. (2011) using various fingerprinting techniques. These are compared with an actual α-pinene SOA mass spectrum obtained by Alfarra et al. (2013) during a chamber experiment.

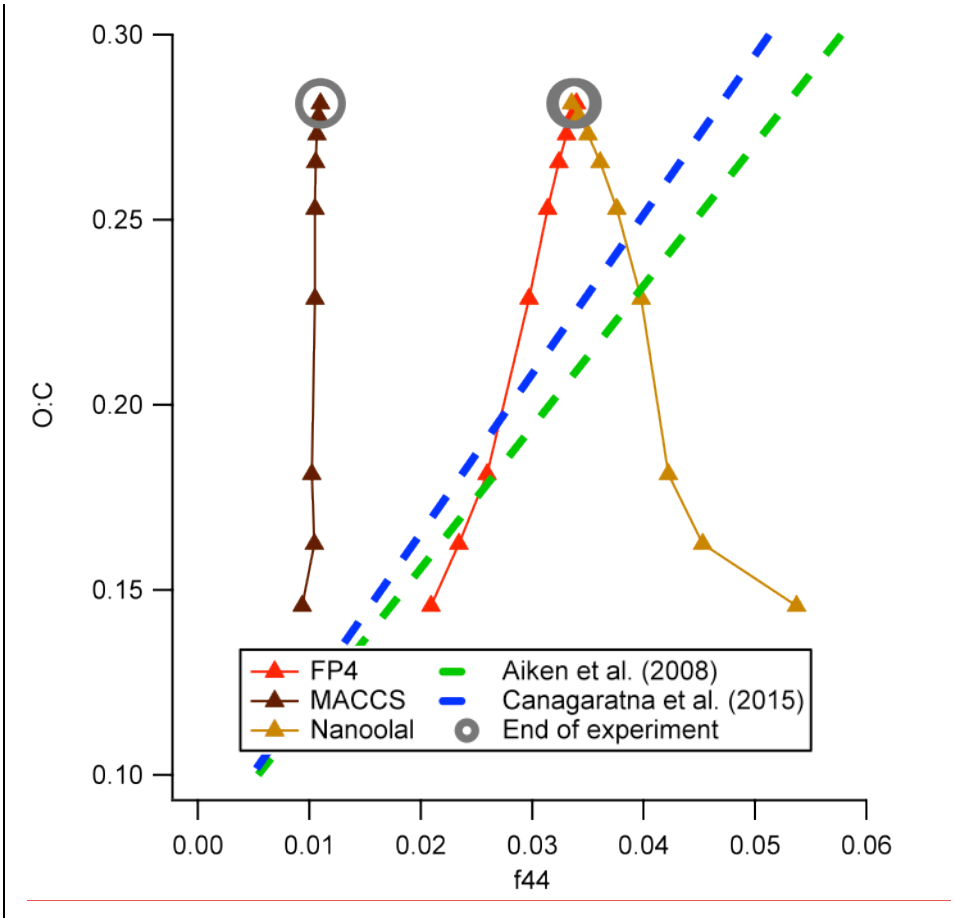

Figure 8 Comparison of O:C ratios and predicted fractional contribution to the AMS m/z 44 channel (f44) for the Valorso et al. (2011) GECKO-A simulation, compared against the regressions performed by Aiken et al. (2008) and Canagaratna et al. (2015). The highlighted points indicate the final points in the simulation.

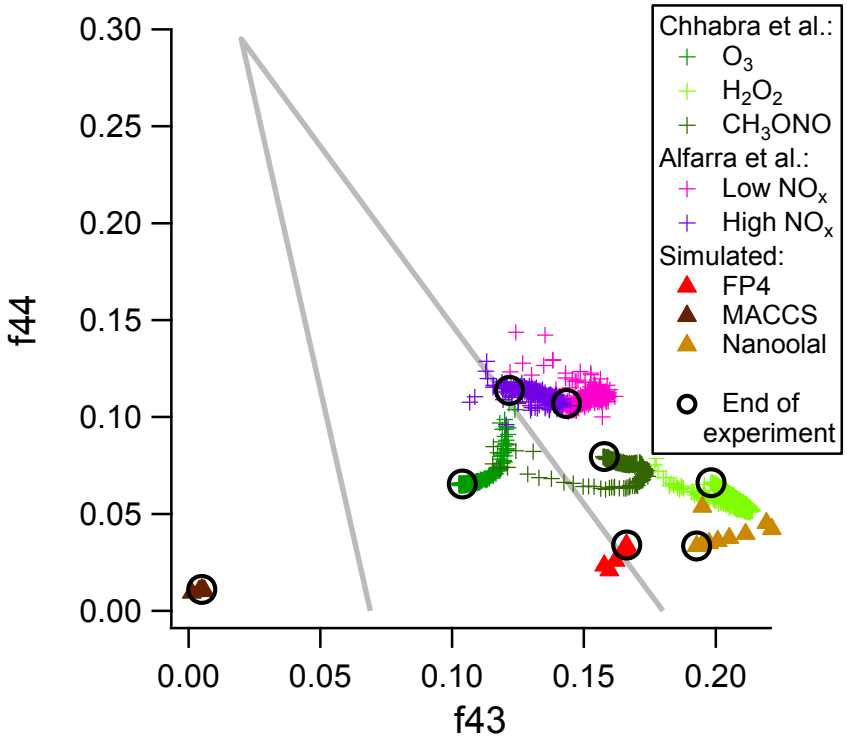

Figure 9 – 'Triangle plot' comparing predicted f44 and f43 values for the Valorso et al. (2011) GECKO-A α-pinene SOA simulation with chamber experiments. The Chhabra et al. (2011) data compares different oxidant systems and is taken from figure 2A of that paper. The chronological final points in each dataset are highlighted.