We would like to thank the reviewer for their recognition of the potential of the approach presented here. In the following we respond to all comments, including detailing some additional work that has been carried out with regards to fingerprint analysis. In the following response we separate and number all distinct comments in order of their appearance in the review, highlighting new text added to the manuscript where appropriate.

1) *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?*

Response: Comparing average performance statistics in section 3.1 at first implies this might be the case. However the comparison with spectra from the Alfarra et al. (2013) paper illustrates the MACCS keys perform poorly. Interrogating the performance from predictions using the MACCS keys for specific compounds illustrates a few problems that might reflect a lack of generality across the MACCS keys. For example, the FP4 keys cycle through systematic functional groupings such as: primary carbon, secondary carbon, tertiary carbon…primary alcohol, secondary alcohol, tertiary alcohol etc. This would lead to a maximum of 320 keys per molecule. MACCS keys on the other hand are almost seemingly designed to capture a random, although extensive, set of features leading to a maximum of 162 features for any given molecule. As we note in the manuscript, it is difficult to find the provenance behind the MACCS keys. However, we have added the following text in section 2, page 5, to try and clarify the issue [new text presented in italics]: '*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 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.'

2). *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?*

Response: We agree this is surprising, especially given the extent of applications to which SVMs are applied. At first we assumed this was down to how the data was normalized prior to training. However, using a maximum/minimum scalar prior to training did not improve performance. There are differences according to which kernel is used. It might be true that dimension reduction procedures, such as PCA, might improve performance. With this in mind, we have conducted tests on using PCA prior to training, using the combined set of fingerprints as requested in point '6' addressed shortly. Based on these results we have added an additional table [table 3] demonstrating the effect of dimension reduction procedures on the performance of all methods, using the combined fingerprint approach:

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.

We have also added the following text to section 3.2 [new text presented in italics], which is renamed to: 3.2 Training to a subset, variable selection **and dimension reductions**. '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 dimension 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 was varied 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.*'

We cannot say with any certainty what the true cause of variability within each regression technique is. Ultimately, we feel this proof of concept study needs building on with appropriate laboratory data before further quantification of dependencies would be possible. Whilst we state the rationale in the original manuscript, we have added the following text in section 4 to re-iterate this: '*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.*'

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

Response: Using the cosine angle between spectra as a measure of good fit, parameters for each method, where required, are cycled until the most effective combination were found. These parameter ranges are presented in the code release and are specific to each algorithm,.

4) *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 under predicted, but this seems a bit buried in the presentation.*

Response: There are indeed other metrics we could have employed to measure distance between mass spectra, however we considered cosine to be the most appropriate. Firstly, because our aim is to replicate the AMS instrument response function, which can be modelled as a linear addition of multiple component mass spectra, we reason that it would make the most sense to use a metric that places linear weight on the peaks' relative intensities. Secondly, while a different metric may place a relatively greater weight on intermediate peaks (thus ensuring a more general agreement over a larger number of peaks), we would have to take care not to also unduly weight the minor peaks, which can be problematic. As such, an element of subjectivity would have been introduced in the choice of algorithm, which in itself would require more testing. It is possible that there is a better closeness metric that could be tested as part of future work and this would be easily testable within the STRAPS framework, however see that as outside the scope of this particular paper. Concerning the comparison between f43 and f44, this refers to the specific comparison between the GECKO-A run and roughly comparable chamber experiments, however we must stress that this test was only to demonstrate proof-of-concept and not perform a systematic comparison to assess the performance. We merely show that the values produced for these two common AMS metrics are plausible in magnitude. For this to be done properly, a chemical model run matched to the exact chamber system should be performed with a state-of-the-art model; this will form part of future work and a full, systematic comparison of peak magnitudes will be performed there.

5) *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)?*

Response: The methodology presented here is based on predicting a response for each channel, and then predicting the peak height for each channel. Each m/z therefore has its own model and there is not dependency on whether 100, 150 or 300 m/z's are chosen. There is no penalty to predicting the high m/z peaks, as these generally represent a low mass fraction and contribute little to the cosine of the comparisons. However, there will be a tangible disadvantage to operating on a reduced dataset because the data reduction in itself will inherently remove information that is possibly of value for training, so there is a very real risk of an inferior training.

6) *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)?*

Response: This is a good point, and we have conducted additional simulations to investigate this. It is worth noting the initial aim of the paper was to illustrate the use of 'standard' fingerprint libraries, as they exist as distinct developments. As noted in the manuscript, ideally we would like to take this proof of concept work forward by constructing a library of keys that better represents the mechanism of fragmentation within the AMS. It might be that converting general rules of EI fragmentation would be a useful starting point. Tables 1-2 now includes median cosine angles from each regression technique when combining all keys into one fingerprint:

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.

We have also added the following text to section 3.1, Page6 [new text in italic]: '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.'

Followed by: '*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*.'

This is now combined with the request presented earlier to assess the role of dimension reductions, using PCA, leading to a new table [3] and subsequent text presented in response to point 2. We also add the following text to the final paragraph in the abstract [new text in italic]:' 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.'

Whilst these new simulations add an interesting angle, we still need more experimental data to resolve any issues with over or under fitting that might occur using our limited, and yet, somewhat disparate set of compounds in the present training database. We feel this is one reason the MACCS keys perform so poorly when methods are applied to the outputs of Valorso et al (2011), in that there are specific keys that are leading to over fitting to the training dataset.

7) *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.*

Response: This has now been added to the figure caption. Concerning the comparison between f43 and f44, this refers to the specific comparison between the GECKO-A run and roughly comparable chamber experiments, however we must stress that this test was only to demonstrate proof-of-concept and not perform a systematic comparison to assess the performance. We merely show that the values produced for these two common AMS metrics are plausible in magnitude. For this to be done properly, a chemical model run matched to the exact chamber system should be performed with a state-of-the-art model; this will form part of future work and a full, systematic comparison of peak magnitudes will be performed there.