We would like to thank the reviewers for their comments. We apologize for the extended time between reviews and revision but this was due to incorporating the important extra work requested in the revised manuscript. We have made the changes suggested and responded to comments in a point by point format below. We believe this is a stronger paper as a result and thank the reviewers for their work.

Reviewer #2

As an overview comment we repeat the introductory review text here. Each point made is repeated specifically addressed below.

The major shortfall of this paper is that the authors neither explain the details of the machine learning approach fully nor do they fully engage with the aerosol classification results, leaving the reader somewhat confused in both realms.

In addition, the authors do not attempt to address the performance of their approach in terms of time or give information about how applicable it would be to ambient data sets where particles would not necessarily be of such distinct types.

Finally, they given no metrics for success – how good is good enough performance for this approach? How good are other methods, compared to that presented here?
I would recommend that this paper be significantly revised, in such a way that a) the machine learning approach can be fully described and choices made justified with data, and b) the aerosol particle classification results can be fully examined and compared to other methods.

Specific Comments:
1. The paper reads as if it was written by two separate people, one for the algorithm discussion and one for the aerosol particle classification discussion. This should be addressed as a final version (or versions) is developed. For example, on p. 4, the transition between lines 12 and 13 is abrupt and jarring.

We regret the paper seemed disjoint and believe this is partially from the fact the we are using a new technique on a more traditional dataset. It was not, as the reviewer suggested, written by two different people and knit together. We have gone through and tried to streamline the flow of the paper by removing the less pertinent details surrounding the algorithm, including your suggestion of simplifying the discussion of confusion matrices – please see the full track changes version. We are attempting to characterize two distinct topics as they related to the paper: The details surrounding training and applying a random forest as well as aerosol populations in the content of mass spectrometry.

To directly address this example, 4:12-17 now reads:

"To pick up on these minor yet important compositional differences, robust and generalizable analysis techniques are critical. We show that supervised training with random forests can differentiate aerosols in SPMS data more accurately than simpler approaches."

2. The authors refer to "volatile" components of aerosol particles multiple times in the paper (first p. 3, line 13). I believe they mean semi-volatile, or at least "more volatile" than other components. Volatile species would not be expected to be found in particles.

The reviewer is correct and we have replaced the term volatile with semi-volatile.

3. In section 2.2, where the training data set is introduced, the authors need to discuss the applicability of this dataset to any "real" experiment. Would these particles be a good representation of ambient particles, for example?

We have addressed your question by including the following.
6: 6-11

> "The choice of supervised or unsupervised machine learning will depend on the researcher's use-case, and each method has unique advantages and disadvantages. We note a limitation of the random forest approach - and for supervised learning in general - is the inability to classify aerosol types outside of the training set. The ability of a random forest to characterize ambient atmospheric datasets, therefore, will strongly depend on which aerosols are contained within the training set."

Although it is feasible that unseen aerosol types will be assigned to the most chemically-similar label, supervised models are tuned to only make predictions on labels in the training set. The error statistics cannot be fully quantified for datasets with unknown aerosol types, so the model may not conform to the determined generalization error. In general, more particle types lead to a more generalizable classifier with better quantifiable error statistics. A study looking chiefly at atmospheric spectra would benefit from adding additional aerosol types and augmenting the analysis with existing methods such as clustering, which are designed to handle unlabeled data.

4. In the discussion of the data presented in Table 2, the authors state that the columns labeled "broad" are applicable to the categorization of the particles when they are lumped together into broad chemically-similar categories. These categories should be defined in this context (and not just in the context of the confusion matrices), presumably in Table 1. Any interpretation of the differences should be discussed in the results section. In addition, the paragraph on p. 11 about the 59+ ion observed in some samples (which ones?) should be moved into the results section. Finally, is it certain that 59+ is Co rather than an organic contaminant?

In this case, the contaminant is almost certainly Co^{+} originating from tungsten carbide grinder used to process some of the dust. A typical spectrum that shows the nature of the contaminant is shown below. The spectrum has markers that correspond to Co^{+}, W^{+2} and W^{+} but no obvious organic markers. Alternate assignments for m/z +59 are possible (and we present them in Table

2), but a prominent m/z +59 peak in this dataset is always associated with tungsten carbide contamination shown below.

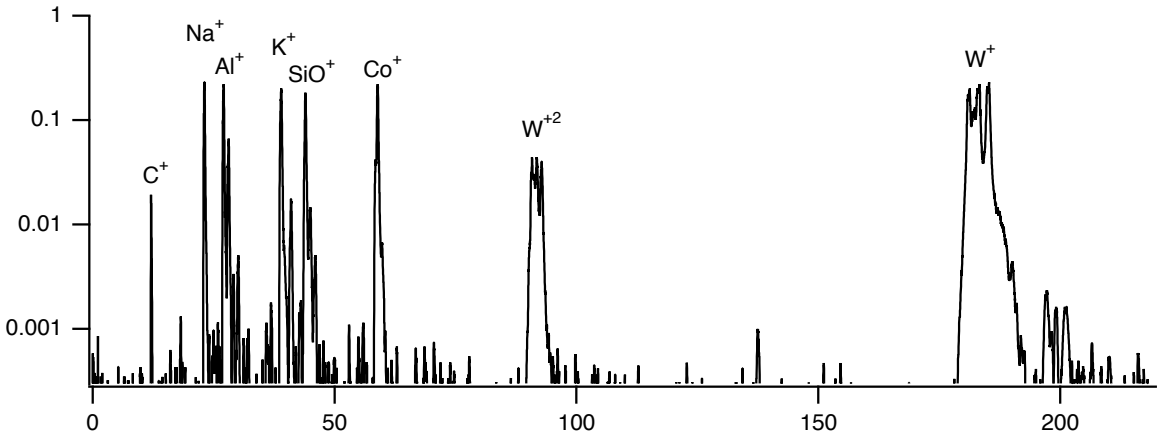


Table 1 has been color-coded to make the broad category definitions more clear. Additionally we move the paragraph into the discussion as requested and expand:

16:22 – 17:5

> "It is noteworthy that while most of the features are logical differentiators of the aerosol types investigated in FIN01 there were also surprises. One example is 59+ (cobalt), determined to be one of the most important features for differentiation. Further investigation determined this material was associated with tungsten carbide contaminant from dry powder dispersion equipment used on some samples. The contamination affected feldspar samples used during the second half of the AIDA measurements in particular."

5. The authors provide no information about the average mass spectra of the individual particle types and the variability within "identical" particles or between particle types. This would seem to be an important parameter in determining how well the algorithm can do to separate them. Based on the two peaks shown in Figure 1, this is an important factor.

To address your point, we have compared the method to a simple classifier which assigns unknown aerosols to the nearest class mean vector using the Euclidean distance metric. This answers your previous question of "how good is good enough" and provides a baseline that directly depends on the distance of aerosol mean vectors in feature space and variability of individual aerosols within each class. Figure 4 is updated to show the results of such a classifier. A couple of paragraphs have been included to summarize these results.
13:10 - 20 now reads:

"To access relative model performance, we contrast the results with a simple classifier that compares unseen aerosols to a set of class mean vectors. Using the Euclidean distance metric, the unknown aerosol is assigned to the nearest class. This simple baseline classifier helps put results in the context of machine learning techniques that rely on distance-based metrics such as k-means and hierarchical clustering. K-means clustering attempts to divide the data points into k distinct clusters, representing spectra as vectors. Using Euclidean distance, the standard algorithm assigns points to centroids, or clusters, which are essentially mean vectors representing the average of all points in the cluster. Assuming perfect convergence of k-means clustering, where k is the number of aerosol classes, each cluster represents the mean of aerosol in that class. The random forest results below demonstrate many areas of improvement over the simple classifier."

16:8 - 20 now reads

"*Across all categories, the random forest shows improvements over the Euclidean classifier in terms of both accuracy and precision. Figure 4 directly compares confusion matrices for the two methods, revealing overall accuracy improvements of at least 20%. The largest improvements are in the fertile soil and other category, where accuracy rises between 20% and 39% with the random forest. Computing the full confusion matrix for the Euclidean technique (as in figure 3) reveals similar results, with far more frequent mislabeling between fertile soils as well as coated/uncoated particles than our approach. These results reinforce the fact that chemically-similar aerosols which overlap in feature space will often be grouped together when using a single, distance-based classifier. The improvement from random forests is likely a result of a) the ensemble approach, which is known to produce better generalizability than single classifiers and b) the tendency of aerosols with similar chemical properties and atmospheric effect to appear mathematically distinct with a distance metric."*

6. The discussion of confusion matrices was confusing. Essentially, these matrices represent normalized counts of the sorting of known particles into the available classes.
This can be stated much more cleanly than the three page description provided on pp. 12 – 14. This is another example of a section that is trying to be both an algorithm and a particle chemistry paper, and not mixing the two effectively.

The section on confusion matrices has heavy revised and simplified to focus on aspects of the matrix that are directly used for the paper.

14: 2-4 reads:

"*A confusion matrix captures misclassification tendencies by pair-wise matching the model prediction with the true aerosol type or broad category [Powers, 2007], and can be understood as a contingency table matching model predictions to true labels."*

7. Figure 5 and the discussion of the "blind" test in Section 3.2 are key to the goals of this paper, but are confusing in their presentation. Regarding the text, why do the authors not know the number of particle types that were used in the challenge (p. 15, line 2 "3-4 aerosol types . .

.were aerosolized")? How well can the results be evaluated if the test conditions are not known?...

The purpose of the blind experiment was to determine the capability of each mass spectrometer to determine the number of particle types and their composition; a situation deemed similar to the challenge of atmospheric sampling. This is now explicitly stated in the text at 17:10 "As part of the FIN01 workshop, it was known that an unknown number of aerosol types from Table 1 were aerosolized into the ADIA chamber at unknown size and relative concentration." We realize the wording in the original sentence implied an unknown test but it was meant to indicate the participants were not aware.

b.) The authors describe a probabilistic correction for the mis-labeling that they observe in their confusion matrices (p. 15, lines 14 – 18), and say that the results "better represent the underlying aerosol population." (p. 15, lines 17 – 18), but they don't provide the data to evaluate this claim.

- The proposed probabilistic correction leads to insignificant changes in the final predictions as the computed fractional difference are small relative a) misclassifications between labels and b) uncertainties from having an unseen label. Furthermore, there is no guarantee the blind dataset will conform to the same mislabeling tendencies, as you have mentioned. We have removed the description from the paper and reapplied the method without the correction.

c.) The data presented in Figure 5 do not make the case that the authors are trying to make. While the two models (positive and negative) show relatively good agreement with each other, the representation of the particles introduced into the chamber is poor. The authors show the breakdown of soot, SOA, and mineral particles introduced in Figure 5, state that the soot particles are too small to see with their instrument, and then compare against the soot-containing dataset anyway. If the pie that represents "Aerosols Reported by AIDA" were renormalized to include only observable particles in this experiment, SOA would represent 44% of the pie and mineral would represent 56% – assuming that the "Aerosols Reported by AIDA" pie is also representing number of particles, rather than mass of introduced particles (this is not stated). If this pie represents something other than number, there is no comparison to the blind sample possible in this figure, only a comparison between the two models.

- The aerosols reported by AIDA are number concentrations, but there are several considerations to account for when considering the labels reported by our classifier in the blind dataset.
    - The aerosols reported by AIDA do not account for PALMS transmission efficiency, which depends on the size and aerodynamic properties of aerosols. For example, particles larger than 1000nm are over-reported by the classifier due to increased PALMS efficiency in that range. We now note this inherent limitation.

- During the course of the experiment, we expect (and observed) the mineral dust and SOA to coagulate. Since aerosol types were reported by AIDA before particles enter the chamber, it is not possible to quantify exactly what fraction of the particles picked up an SOA coating. Additionally there is the possibility of effectively producing a particle type not in the training set, depending on the exact mineralogy of the mineral dust used by AIDA. While it is known a mineral dust was included in the chamber, the exact composition of the dust is not know. The mineral may either contain a specific component or a soil dust. While our training set contains K-Feldspar coated with SOA, a different type of SOA-coated mineral dust will appear unique to the model. Because the generalization performance of supervised classifiers is ill-defined for particles not included in the training set, this could lead to performance that is not captured by the confusion matrices. Given the experimental uncertainties from transmission efficiency and coagulation, as well as the model uncertainties highlighted in the confusion matrices, we believe the results reveal skill in using random forests to pick out distinct aerosols. In future studies, uncertainties can be reduced by adding additional particle labels or accounting for transmission efficiency, but coagulation will likely remain an inherent uncertainty.

These are stated at the end of our results section as well as the caption for Figure 5. The caption in Figure 5 has been updated to mention transmission efficiency as an uncertainty.

35: 6-10 now reads

"Notes (1) the soot in the blind mixture was known to be below the instrument detection limit and therefore is not expected to be found in the data, (2) coagulation of SOA and mineral dust, which occurred after aerosol input to the chamber, was often categorized as mixed mineral and organic particles or fertile soils (i.e., mixtures of mineral and organic components) considered in the training data set, (3) the aerosols types reported by AIDA do not account for PALMS transmission efficiency (see text for details). "

Technical Corrections:

1. P. 10, line 17: remove the word "rows" from the line.

   This has been corrected.

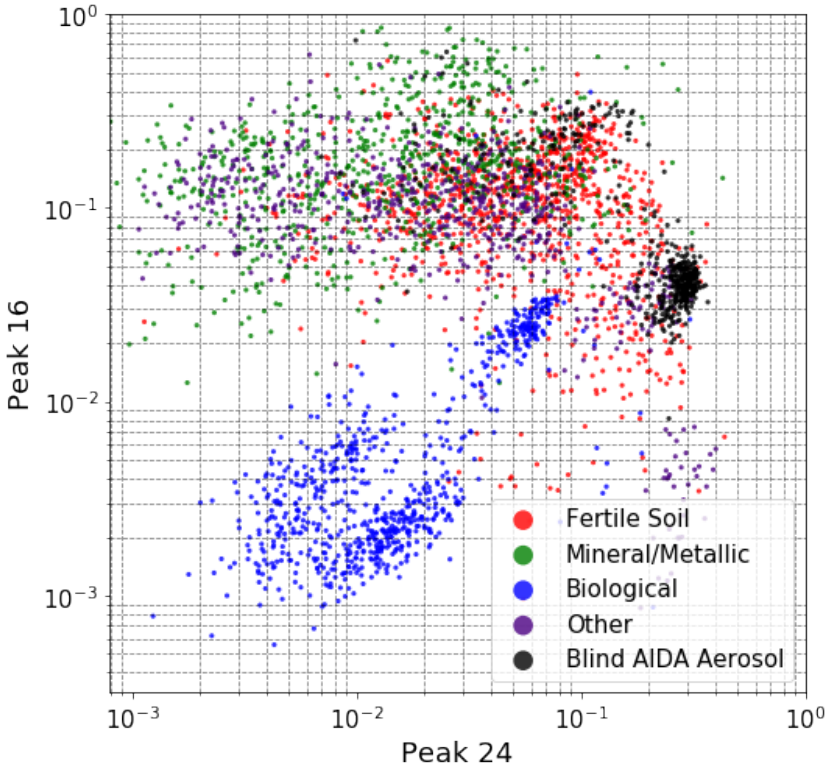
2. P.15, line 3: "AIDA" is written as "ADIA."

   This has been corrected.

3. Figure 1 is very difficult to read. The black points for the "Blind AIDA Aerosol" are only visible on the right-most part of the graph, in the region of (0.3, 0.04) and the similar colors are hard to differentiate. Consider a figure like this that is broken out into the broad categories of particle types.

As suggested, we created a similar scatter plot using only broad categories. We still find it difficult to pick out each of the particle types in the region in (0.3, 0.4), even when plotting only small subsets of the training set. This is a result of significant overlapping of aerosol types in the region, and difficult to alleviate. Nevertheless, the aim of the figure is to demonstrate that some types overlap significantly, while others such as SOA form distinct clusters. Since both versions

of the plot demonstrate this to a similar degree, we have decided to leave the original figure in the paper.



4. Figure 5, negative model, the small pie includes 5 wedges but only 4 labels.
This has been corrected.