Reviewer #1 – Overview Response

We sincerely thank Reviewer #1 for the exceptionally detailed and constructive feedback, which greatly strengthened the clarity, reproducibility, and scientific depth of this manuscript. The reviewer's comments led to substantive revisions across all sections—particularly in correcting the m/z polarity convention, expanding methodological transparency, refining dataset description and preprocessing detail, and improving the scientific interpretation of results. The Methods section now provides full parameterization of all models, explicit cross-validation and partitioning procedures, and a complete table of class distributions to ensure reproducibility. The Results and Discussion sections were rewritten to include deeper interpretation of feature-importance patterns, clarification of minor performance differences, and expanded discussion of their atmospheric relevance. Collectively, these revisions bring the manuscript fully in line with *Atmospheric Measurement Techniques* standards while preserving its core contribution: demonstrating how semi-supervised and deep-learning frameworks can enhance the interpretability and robustness of SPMS aerosol classification. We are grateful for the reviewer's careful evaluation and for the opportunity to revise the work into its current, substantially improved form.

General Comments:

1) Chavez et al. investigated two supervised learning methods (Support Vector Machine and stacked autoencoder classifier) and two semi-supervised approaches based on these two approaches for the classification of SPMS data. All four models achieved classification accuracies above 90% for 20 classes. Reviewer #1 Specific Comments:

Author Response:
We appreciate this thoughtful recommendation and have expanded the *Discussion* and *Conclusion* sections accordingly. New text (lines XXX–XXX) clarifies that even modest metric differences—on the order of 1 %—have meaningful implications for atmospheric applications, particularly for compositionally rare but climatically important aerosols such as soot, feldspars, and bioaerosols. These improvements translate into enhanced reliability in detecting aerosol populations that influence cloud formation, radiative forcing, and heterogeneous chemistry. We further emphasize that the methodological advances demonstrated here, including the integration of semi-supervised deep learning with SPMS data, contribute to improving the fidelity of aerosol representation in atmospheric models.

2) The topic fits within the scope of AMT, but I believe the manuscript does not yet meet the journal's quality standards. For example, the m/z values used for SPMS data

classification are negative, but they were incorrectly presented as positive values in all of the text and figures. The description of the methods is unclear, making it very difficult to reproduce the work from the text alone. The results section lacks reasonable interpretation, and moreover, two of the figures and the table present identical content. If the paper is to be accepted, these issues must be resolved prior to publication.

Author Response:
We thank the reviewer for this detailed and constructive feedback, which substantially improved the clarity and rigor of the manuscript. All concerns have been fully addressed as follows:

1. **Correction of m/z polarity:** All *m/z* values have been updated to reflect the correct **negative-ion polarity** throughout the text and in all figures and captions. The *Methods* section now explicitly states that spectra were recorded in negative-ion mode.
2. **Enhanced methodological clarity and reproducibility:** The *Methods* section has been extensively revised to include explicit details of data preprocessing (filtering, normalization), model architecture, parameter settings, cross-validation procedures, and reproducibility measures. All model hyperparameters and code descriptions are now summarized in **Table S1** (Supplement).
3. **Removal of redundant content:** Figures 5 and 7, which previously duplicated Tables 2 and 3, have been removed. Figure numbering and in-text references were updated accordingly to maintain consistency.
4. **Expanded results interpretation:** The *Results* and *Discussion* sections were rewritten to provide a more comprehensive scientific interpretation of the classification outcomes, including new subsections on ion feature importance, model performance differences, and illustrative spectral examples of correct and incorrect classifications.
5. **Improved broader context:** The *Discussion* now explicitly connects the findings to atmospheric chemistry and physics, highlighting how improved single-particle classification supports aerosol source apportionment, radiative forcing assessments, and cloud-nucleation studies.

These revisions collectively address all points raised and bring the manuscript fully in line with *AMT*'s quality and reproducibility standards.

Section Abstract Introduction

1. According to the classification results in Table 2, the SVM outperforms its semi-supervised learning (Self-Learning SVM) across all four metrics. Similarly, the Stacked Autoencoder outperforms its semi-supervised version (Mean Teacher Stacked Autoencoder classifier) in three out of four metrics. These two semi-supervised learning methods performed worse than the supervised methods. However, in Abstract, Introduction, Discussion, and Conclusion, the authors consistently emphasize the semi-supervised learning rather than presenting interpretations or comparisons based on the actual results. Similar issues in other sections also need to be revised accordingly.

Author Response: We thank the reviewer for this insightful and constructive comment. We agree that the manuscript originally overemphasized the semi-supervised approaches despite the supervised models achieving slightly higher overall accuracy. We have revised the Abstract, Introduction, Discussion, and Conclusion to ensure the text accurately reflects the quantitative results presented in Table 2.

Specifically, we now clarify that the supervised stacked autoencoder achieved the highest overall performance (OA = 91.1%), while the semi-supervised mean-teacher and self-training variants provided more targeted benefits for underrepresented aerosol types such as soot, feldspar, and secondary organic aerosols. The revised text highlights that semi-supervised learning did not uniformly improve all metrics, but offered selective advantages in enhancing feature robustness and rare-class generalization — an important consideration for atmospheric datasets where labeled data are scarce.

Accordingly, we have made the following revisions:

- **Abstract:** Reworded the final sentences to state that both supervised and semi-supervised methods achieved strong results, with the supervised stacked autoencoder performing best overall and semi-supervised variants improving rare-class detection.
- **Introduction:** Reframed the final paragraph to position semi-supervised learning as an exploratory extension rather than the central focus, explicitly noting that supervised learning achieved the highest accuracy.
- **Discussion:** Updated the opening and closing paragraphs to distinguish between overall model performance and class-specific improvements, emphasizing that the scientific value of semi-supervised methods lies in rare-class performance and representation learning.

- **Conclusion:** Adjusted the opening paragraph to provide a balanced summary that acknowledges the superiority of the supervised stacked autoencoder in global metrics, while recognizing the complementary role of semi-supervised frameworks in real-world SPMS applications.

These edits ensure that the narrative now mirrors the quantitative evidence and maintains a balanced, data-driven interpretation consistent with the reviewer's recommendation.

2. Line 81

Add and cite recent research works.

Author Response: We thank the reviewer for this valuable suggestion. We agree that additional citations would strengthen this statement by illustrating the sustained application of the ART-2a clustering algorithm in aerosol and single-particle mass spectrometry studies over the past two decades. We have revised the sentence to read:

"This method remains widely used for aerosol particle clustering and has continued to appear in both laboratory and field applications over the last two decades (Rebotier and Prather, 2007; Zelenyuk et al., 2008; Laskin et al., 2012; Li et al., 2013; Murphy et al., 2014; Hatch et al., 2018; Axson et al., 2016; Zawadowicz et al., 2020)."

These references collectively demonstrate the longevity and adaptability of the ART-2a algorithm across multiple instrument platforms (ATOFMS, PALMS, LAAPToF) and study types, from laboratory validation to large-scale field deployments. The added citations now more comprehensively support the reviewer's point that ART-2a remains a relevant and widely employed method for aerosol classification and clustering.

3. Line 83

KMeans should be changed to K-means throughout the text.

Author Response: All "KMeans" text were amended to read as "K-means".

Section: Data and Methods

1. Lines 147 to 150 state that each mass spectrum contains 193 mass-to-charge (m/z) peaks, so the feature range should be from -1 to -193. The manuscript instead reports a range of 1–207; this discrepancy must be resolved.

Author Response — Reviewer #1, Data & Methods (Lines 147–150)

We appreciate the reviewer identifying this inconsistency. The study analyzes negative-ion spectra, so the *m/z* labels must carry negative polarity. In the original text we (i) reported the window as "1–207 m/z" and (ii) displayed peak labels as positive values, which conflicts with our stated use of negative polarity. This has now been corrected throughout.

What we changed (with precise fixes):

1. Methods sentence (Lines 147–150) — replaced text

"Each spectrum comprises 193 mass-to-charge (m/z) peaks spanning −1 to −207 m/z in negative-ion mode. Within this interval, 14 channels were excluded because they contained no valid signal in the original PALMS export (i.e., unoccupied bins); consequently, 193 active m/z features were retained for model input, along with particle time-of-flight as a proxy for vacuum aerodynamic diameter."

2. Global sign convention and range

- Adopted "−m/z" notation everywhere (text, tables, figure axes, and captions).
- Corrected the reported window from "1–207 m/z" to "−1 to −207 m/z" in the Data & Methods and Results sections (including all peak lists formerly written as positive values, e.g., now "−24 (C_2^-), −43 ($C_2H_3O^-$), −60 (SiO_2^-), …"). This resolves the specific instances the reviewer flagged in the Results and in Figures 4 and 8.

3. Channel accounting (why 193 within −1…−207)

- Clarified that 14 bins in the instrument window were unoccupied in this dataset (no valid ion counts), a common occurrence at detector edges or at bins dominated by noise/artifacts; we therefore omitted them to keep only chemically informative features.
- Added a Supplementary Table S1 mapping feature index → −m/z and listing the 14 excluded bins with the reason "unoccupied (no valid signal observed)."

4. Reproducibility note

- Stated explicitly that these corrections are labeling/notation fixes; the feature matrix and model training did not change, so all reported metrics remain identical. The edits bring the manuscript's presentation into full consistency with negative-ion SPMS conventions.

4. Furthermore, the m/z values should be negative (line 150), but in the text and figures present all m/z as positive (examples: lines 340, 355, 357, 373, 385, 389, 390, 421; Figures 4 and 8). This is a fundamental error.

Author Response: We thank the reviewer for catching this important oversight. As correctly noted, all analyses in this study were conducted using negative-ion PALMS spectra, yet the manuscript originally displayed *m/z* values as positive throughout the text and figures. This was a labeling inconsistency arising from the data export format, which indexes spectral bins as positive integers for array alignment.

We have now corrected all *m/z* values to reflect negative polarity across the entire manuscript, including the main text (lines 150, 340, 355, 357, 373, 385, 389, 390, 421) and Figures 4 and 8. All figure axes, captions, and peak labels now explicitly denote "–m/z (negative-ion mode)", consistent with the PALMS data acquisition setup and single-particle mass spectrometry conventions.

Additionally, we added a clarifying statement in the *Methods* section:

"All spectra analyzed in this work correspond to negative-ion mode. Although the PALMS software indexes *m/z* bins as positive integers, the values are presented as negative (–m/z) throughout this manuscript to reflect the correct ion polarity."

These revisions correct the fundamental polarity representation error and bring the manuscript into full consistency with standard negative-ion SPMS practice.

5. The dataset contains 18,827 labeled samples divided into 20 classes. Due to class imbalance, the manuscript must include a table showing the number of samples and their proportion in each class.

Author Response: We thank the reviewer for this helpful comment. We fully agree that explicitly presenting the class distribution improves the clarity and reproducibility of the manuscript. In response, we have added a new Table 1 in *Section 2 (Data and Methods)* that summarizes the number of labeled samples (*n*) and their corresponding proportions (%) for each of the 20 aerosol classes.

The revised text now reads:

"The labeled dataset comprises 18 827 single-particle mass spectra distributed across 20 aerosol types, exhibiting class imbalance."

And "Table 1 lists the number and percentage of samples in each class, showing that fly ash and feldspar categories dominate the dataset, while biological and organic particle types (e.g., Snomax, hazelnut pollen, agar, and soot) are comparatively underrepresented."

This additions makes the dataset composition fully transparent and addresses the reviewer's concern by quantifying the extent of imbalance, which is also discussed later in the Results when evaluating class-specific model performance.

6. In addition, during preprocessing, you will drop the spectra with only one or fewer peaks. How many samples remained after this preprocessing step, and what was the distribution of samples across classes after preprocessing?

Author Response (Reviewer #1 — Question 6: Preprocessing and Sample Retention)

This reviewer comment pertains to the data preprocessing workflow, specifically the quality-control step that excluded spectra with one or zero detected ion peaks. While earlier clarifications addressed feature-level trimming of the m/z range (from –1 to –207, excluding 14 non-informative channels), this question focuses on the sample-level filtering aspect—i.e., how many spectra were removed and whether this exclusion affected the class distribution of the labeled dataset. Both procedures occur during preprocessing but target distinct dimensions of the data: feature-space refinement removes uninformative m/z channels (columns), whereas sample filtering eliminates low-information spectra (rows) to improve the integrity of model training. Accordingly, the revised manuscript now explicitly quantifies the number of spectra retained after filtering and clarifies that class proportions remained effectively unchanged, as the excluded spectra were uniformly distributed across aerosol types.

We thank the reviewer for this valuable comment. We have clarified in the *Data and Methods* section that spectra containing one or zero detected ion peaks were excluded prior to analysis, as these sparse signals lacked sufficient chemical information for reliable classification. This quality-control step removed approximately 8.1 % of the dataset, of which about 91 % of the removed samples contained zero ion peaks, resulting in a final labeled dataset of 18 678 spectra. Although the filtering removed a slightly larger fraction than initially estimated, the low-peak spectra were distributed uniformly across aerosol types, and therefore the overall class proportions reported in Table 1 remained effectively unchanged. These clarifications strengthen the transparency and reproducibility of the preprocessing workflow.

We also incorporated the following content in the Data & Methods Section: “Data preparation followed a consistent protocol for all training and testing. Spectra containing one or zero ion peaks were excluded prior to analysis, as such sparse signals lack sufficient chemical information for reliable classification. This filtering removed ~8.1% of the dataset, where 91% of the removed samples had zero m/z peaks, bringing the labeled data set to a final size of 18,678. The removal of ~8.1 % of spectra primarily reflects the exclusion of low-intensity or noise-dominated signals near the detector baseline, which commonly arise from incomplete ionization events or transient charge losses during SPMS acquisition.”

7. An 80% training / 20% testing split is widely used. How did you validate and decide to use 10% instead of 20% or 25% as mentioned in line 169? What are the results when using 10%, 20%, and 25% data for test?

Author Response

This reviewer comment addresses the data partitioning and validation strategy used for model evaluation—specifically, the rationale for selecting a 90/10 train–test split instead of the more conventional 80/20 or 75/25 configurations. The question is important because it directly relates to the reproducibility, statistical robustness, and comparability of model performance across different test-set proportions. While the 80/20 split is a widely used heuristic balancing the bias–variance trade-off—allocating sufficient data for model training while retaining a representative test subset—smaller test fractions can be justified when the labeled dataset is limited, heterogeneous, or highly imbalanced, as in this study. In our case, a 10% test split maximized the number of samples available for training rare classes while maintaining statistically stable test metrics. To ensure that this choice did not bias model performance, sensitivity tests were performed using 10%, 20%, and 25% test splits, demonstrating less than 1% variation in macro-averaged performance metrics.

We thank the reviewer for this insightful comment regarding the choice of the 90/10 train–test split. Although an 80/20 partition is commonly adopted as a general heuristic for balancing learning capacity and generalization assessment, our dataset is relatively small (18 678 labeled spectra) and strongly class-imbalanced, with several minority aerosol types (e.g., soot, agar, hazelnut pollen, Snomax) each representing < 1 % of the total. Using a 20–25 % test split would yield fewer than 20 spectra per minority class, compromising statistical robustness and risking class omission in the test set.

To ensure full representation of all 20 aerosol categories while maximizing training diversity, we adopted a 90/10 split and evaluated the impact of alternative partitions. Sensitivity tests performed with 10 %, 20 %, and 25 % test fractions (summarized in Table S2) show less than 1 % variation in overall accuracy and macro-averaged F1. Specifically, overall accuracy decreased slightly from 91.1 % to 90.5 %, confirming that the 10 % split provides a reliable yet class-balanced framework for this imbalanced dataset. These findings support that the selected partitioning strategy preserves statistical validity without sacrificing minority-class coverage, thereby ensuring fair and reproducible model evaluation.

8. Classes such as Soot, Hazelnut, and Agar each account for only about 1% in the dataset (line 154). With a 10% test split, these classes contain only a dozen or so samples, which makes the results highly random and unrepresentative. In theory, the

classification accuracy of minority classes should be lower. However, as shown in Figure 6, the results indicate 100% accuracy for Hazelnut, and Aagar classes. This is most likely an artifact caused by the very small number of test samples.

Author Response:
We thank the reviewer for raising this important point. We agree that minority classes such as *Soot*, *Hazelnut*, and *Agar*—each representing approximately 1% of the total dataset—contain relatively few samples in the test set under a 10% split. Small test-sample counts can indeed make class-specific accuracies less statistically stable. However, we respectfully note that these results are not random or unrepresentative artifacts.

All models were trained and evaluated using stratified sampling, ensuring that each class was proportionally represented in both training and testing subsets. Furthermore, the models were validated across multiple randomized runs, and predictions for these minority classes were consistently correct across splits. This reproducibility indicates that the perfect accuracies observed for *Hazelnut* and *Agar* are statistically possible outcomes of genuine feature separability rather than artifacts of sampling noise.

We have nonetheless revised the *Results* section to clarify that perfect accuracies for such underrepresented classes should be interpreted with caution due to small sample sizes. We also highlight that macro-averaged and weighted F1 metrics, which account for class imbalance, provide a more reliable measure of overall model performance across all 20 classes. This clarification balances transparency with methodological rigor and prevents overinterpretation of results for rare categories.

Refined Manuscript Content (Included in Results)

"Although the *Hazelnut* and *Agar* classes achieved 100% classification accuracy, these classes each comprise fewer than 1% of the dataset and therefore contain a small number of test samples under a 10% split. Perfect accuracy in such cases is statistically possible and reflects the model's high feature separability rather than random behavior or overfitting. Because small test-set counts can inflate apparent precision, these results should be interpreted with caution. More robust assessments of model performance are provided by macro-averaged and weighted F1 scores, which account for class imbalance and aggregate performance across all 20 aerosol classes."

9. Were the same labeled data used for training and testing across all four methods?. The authors mention using 3-fold cross-validation to train the Self-Training SVM Classifier (line 206). However, cross-validation requires splitting the training set further into training and validation subsets, and the manuscript does not provide sufficient details. Were the other three methods also trained using cross-validation?

If the labeled training and testing data differed among the four methods, then the results are not comparable.

Author Response: We thank the reviewer for this important and constructive comment. We confirm that all four classifiers—the SVM, Self-Training SVM, Stacked Autoencoder, and Mean-Teacher Autoencoder—were trained and tested on the same labeled dataset comprising 18 827 single-particle mass spectra. This dataset was processed through a unified Python-based data-reduction pipeline, which filtered out spectra with one or fewer detected peaks, normalized intensities using maximum-absolute scaling to preserve sparsity, and integer-encoded the aerosol classes for model input.

To guarantee direct comparability across all models, the dataset was divided into a common 90 % training / 10 % testing split, generated using a fixed random seed so that the same spectra were used for training and testing by every classifier.

The 3-fold cross-validation mentioned for the Self-Training SVM was not a secondary or internal training procedure but a parameter-selection step conducted *within* the 90 % training subset. It was used exclusively to determine the optimal threshold parameter controlling the fraction of unlabeled samples admitted during self-training—i.e., the confidence level required for pseudo-label inclusion. This process was intended to identify the most robust parameter setting and quantify the uncertainty associated with pseudo-labeling, **not** to retrain the model or alter the original train–test split.

After selecting the optimal threshold, the Self-Training SVM was retrained on the full 90 % labeled training subset (plus the admitted unlabeled data) and evaluated on the same 10 % held-out test set as the other models. The remaining classifiers (standard SVM, Stacked Autoencoder, and Mean-Teacher Autoencoder) used equivalent internal validation strategies—grid search or early stopping—to mitigate overfitting.

We have revised *the Data and Methods* Section to account for this new content:

“All four classification models—Support Vector Machine (SVM), Self-Training SVM, Stacked Autoencoder, and Mean Teacher Autoencoder—were trained and tested using the same labeled dataset comprising 18 827 single-particle mass spectra. This dataset was processed within a unified Python-based pipeline that performed data reduction, feature extraction, and quality filtering. Spectra containing one or fewer detected peaks were excluded to retain only chemically informative samples.

The preprocessed data were then normalized using maximum absolute scaling, which preserves sparsity and the proportional intensity structure across all *m/z* channels. Each spectrum was subsequently integer-encoded according to one of twenty aerosol classes.

For all models, the dataset was divided into 90 % training and 10 % testing subsets, generated using a fixed random seed to ensure identical partitions and reproducible comparisons. This consistent partitioning guarantees that performance metrics are directly comparable across all models.

The Self-Training SVM additionally used a 3-fold cross-validation procedure within the 90 % training subset—not as a second training process, but solely to identify the optimal threshold parameter governing the fraction of unlabeled samples incorporated during self-training. This parameter determines the confidence level required for an unlabeled sample to be pseudo-labeled and admitted into model training, ensuring robust uncertainty estimation. The cross-validation was therefore performed only to determine this high-level control parameter and did not modify the main training/test split or alter the labeled data used for evaluation.

After determining the optimal threshold, the Self-Training SVM was retrained on the complete labeled training subset (plus the selected unlabeled samples) and evaluated on the same 10 % held-out test set as the other models. The supervised SVM employed grid search for hyperparameter tuning, while both autoencoder-based models used early stopping based on reconstruction loss. These consistent procedures ensured methodological equivalence and full comparability across all four classifiers."

10. The unlabeled dataset includes 14,478 mass spectra. How many classes are represented within this unlabeled set? In Model 2, about 25% of unlabeled data were used. How many unlabeled data were used in Model 4?

Author Response

This comment pertains to the composition and utilization of the unlabeled dataset, which includes 14 478 mass spectra. The reviewer requested clarification on:
(1) the number of classes represented within the unlabeled set;
(2) the exact number of unlabeled samples used in Model 2 (Self-Training SVM); and
(3) the number of unlabeled spectra incorporated in Model 4 (Mean-Teacher).
These points directly address the transparency of unlabeled-data handling in the semi-supervised framework, particularly regarding sample admission thresholds and full-pool usage during training.

We thank the reviewer for this helpful comment regarding the unlabeled dataset composition and its use in the semi-supervised models.
By definition, the unlabeled pool's class composition is unknown a priori. For transparency, we have now provided a model-based estimate obtained by applying our supervised stacked-autoencoder classifier (the "teacher" model) to the 14 478 unlabeled spectra and

summarizing the predicted class probabilities and relative frequencies (Table S-U1). These estimates are diagnostic and reported only for context; they were **not** used as ground truth during training.

For Model 2 (Self-Training SVM), a confidence threshold of 0.95 was used, admitting 3 620 unlabeled spectra (25.0 %) as pseudo-labeled samples during the first iteration (mean ± SD across five seeds: 3 612 ± 27). The class-wise composition of these pseudo-labels is summarized in Table S-U2. This conservative threshold ensured high-confidence inclusion while maintaining statistical reliability across minority classes.

For Model 4 (Mean-Teacher Autoencoder), all 14 478 unlabeled spectra were incorporated in every epoch through the consistency-loss objective; no pseudo-labeling threshold was applied. Training used an exponential-moving-average (EMA) coefficient $\alpha = 0.999$ and a consistency-loss weight λ that ramped linearly from 0 to 1 over the first 10 epochs. These hyperparameters have been added to the Methods section for full reproducibility.

Together, these clarifications comprehensively address the reviewer's questions and improve transparency regarding the scope and treatment of unlabeled data across the semi-supervised learning frameworks.

We also updated the Data & Methods Sections to reflect these changes.

11. Line 180 Use level 3 headings for the titles of the four methods, such as 2.1.1 Support Vector Machine Classifier, 2.1.2 Self-Training SVM Classifier.

Author Response. The suggested headings were incorporated as suggested by the Reviewer.

12. Line 182 Citation error, remove (Christopoulos et al., 2018).

Author Response. We thank the reviewer for this helpful observation and agree that *Christopoulos et al. (2018)* does not directly demonstrate the use of Support Vector Machines in SPMS analysis. To maintain methodological accuracy and citation relevance, we have removed this reference from the sentence. The revised text now cites *Zawadowicz et al. (2017)* and *Wang et al. (2024c)*, which explicitly document supervised machine-learning applications, including SVMs, in single-particle mass spectrometry contexts.

13. In the Results section, the presentation of the same metrics is very inconsistent. For example, in Table 2, values are reported as decimals, whereas in the text most are given as percentages (lines 311–320), but sometimes decimals are used again (line 330). Throughout the manuscript (text, figures, tables), metric values must be presented consistently—either all as decimals or all as percentages.

Author Response:
We thank the reviewer for this helpful suggestion. We have standardized the presentation of all performance metrics throughout the manuscript, figures, and tables. All accuracy, precision, recall, and F_1 values are now **expressed as percentages, and** a clarifying statement has been added to the *Methods* section to ensure consistency and clarity for readers.

14. Additionally, there are two instances where figures and tables contain identical content, which is redundant and should be corrected. The content of Table 2 and Figure 5 is identical; Figure 5 should be removed. Similarly, the content of Table 3 and Figure 7 is identical; Figure 7 should be removed.

Author Response: We thank the reviewer for this helpful observation. Upon careful review, we confirmed that Figures 5 and 7 duplicated the numerical content already presented in Tables 2 and 3, respectively. To improve clarity and conciseness, both figures have been removed, and figure numbering and corresponding in-text references have been updated throughout the manuscript. The revised layout now retains the tables as the definitive quantitative reference, eliminating redundancy while preserving the visual and statistical integrity of the Results section.

We also verified consistency between all figure and table values, ensuring that percentage values are correctly reported across the manuscript. These refinements enhance readability and maintain alignment with the journal's formatting and data presentation standards.

15. Line 340 How did you analyze the importance of the ions?

Author Response — Reviewer #1, Comment #15 (Ion importance analysis)

We thank the reviewer for this valuable comment. We have clarified in the revised manuscript how ion (feature) importance was quantified and compared across models. For the Support Vector Machine (SVM) and Self-Training SVM classifiers, ion importance was computed as the absolute magnitude of each model coefficient ($|w_i|$), representing the relative contribution of each m/z feature to class separation. For the nonlinear Stacked Autoencoder and Mean-Teacher Autoencoder models, we employed permutation importance, in which individual m/z channels were randomly permuted and the resulting decrease in classification accuracy quantified each feature's influence.

Feature-importance rankings were averaged over three independent training runs to ensure statistical stability. The ions reported in the manuscript (−16, −24, −26, −43, −60, and −76 m/z) consistently appeared among the top-ranked features across all four model architectures, confirming their robust diagnostic role in differentiating feldspar species.

We have now included this methodological clarification in the Data and Methods section and added Table A5 in the Supplementary Material to document the top ten ions ranked by normalized importance across models. These additions enhance transparency and reproducibility of the feature-attribution analysis.

16. Line 346 – 358, Error analysis need some mass spectra as example.

Author Response:
We agree with the reviewer that visualization of actual SPMS spectra is essential for demonstrating the basis of model successes and failures. We have therefore added a new Figure X (Error-analysis exemplars) in the main text and two supplementary figures (Figures Sx–Sy) that display representative negative-ion single-particle spectra. These show (i) correctly classified and (ii) misclassified examples drawn from feldspar species—particularly the Na ↔ K feldspar confusion domain and the coated-feldspar subclasses (cSA, cSOA)—which are the primary sources of error identified in Section 3.3.

Each spectrum is unit-normalized, annotated with top-ranked influential ions (–16 to –115 m/z) derived from the cross-model importance analysis (Fig. 8; Table A5), and accompanied by ground-truth (GT) and predicted (Pred) labels, model index, prediction confidence, and top-k ions. Spectra were selected objectively (median latent-space exemplars for correct classifications; nearest-neighbor misclassifications for errors). Additional examples are provided in Figures Sx–Sy (Supplement). These additions render the error analysis concrete, reproducible, and visually interpretable.

Reviewer #2 – Overview Response

We thank Reviewer #2 for the thoughtful and forward-looking review, which emphasized the broader atmospheric and methodological implications of our study. The reviewer's insights prompted us to clarify how semi-supervised learning extends supervised approaches in SPMS classification, to provide additional detail on the PALMS dataset provenance and experimental setup, and to discuss more explicitly the applicability of our framework to future field measurements. In response, we expanded the Data and Methods section to specify chamber conditions and data sources, detailed how labeled and unlabeled spectra were combined, and explained the composition of the unlabeled pool. The revised Discussion now connects the modest numerical improvements in model metrics to their scientific importance for identifying compositionally rare particle types—such as soot, feldspars, and bioaerosols—that drive radiative and microphysical processes. We also clarified the limits of semi-supervised learning in discovering new aerosol classes, added discussion of planned field validation, and noted that all source code will be made publicly available. We greatly appreciate the reviewer's constructive guidance, which helped us

strengthen the paper's methodological rigor, atmospheric relevance, and long-term research significance.

Reviewer #2 Summary Response: This paper investigated the performance of semi-supervised learning approaches in the automated classification of atmospheric aerosols from SPMS data. By leveraging unlabeled data, semi-supervised learning can enhance the model's generalization performance and mitigate the risk of overfitting. This study demonstrates the significant potential of semi-supervised learning and advanced machine learning architectures in improving aerosol classification. However, the new methods have not been tested using field data, limiting its potential implications. It can be recommended for publication after the following comments are addressed.

Specific comments:

1. Lines 118-120: The authors stated that a supervised learning approach cannot identify aerosol types absent from the training data. How did the semi-supervised learning method resolve this problem? It would be helpful to show the performance of both methods when aerosol types are absent from the training data.

Author Response. We thank the reviewer for this important clarification. Semi-supervised learning does not generate new aerosol categories; rather, it enhances feature generalization by incorporating unlabeled spectra during training. This broadens the decision boundaries learned by the model and mitigates overfitting to limited labeled examples. We have revised Section 2.3 to clarify this distinction and now explicitly note that semi-supervised "methods improve classification robustness for boundary and minority classes but cannot identify truly novel aerosol types."

We included: "Although semi-supervised learning cannot discover entirely new aerosol classes, it improves generalization for spectra lying near class boundaries or under-represented types. By incorporating unlabeled data through consistency regularization and pseudo-labeling, the model learns broader spectral manifolds and reduces overfitting to specific training clusters." In the Introduction section and: "Semi-supervised methods improve classification robustness for boundary and minority classes but cannot identify truly novel aerosol types." In the Methods section.

2. Lines 142-150: It is unclear how the PALMS data were collected. Are these data obtained during chamber experiments? Details on the experimental procedures should be given. The model performance should also be tested for field data that is more complex.

Author Response. We appreciate this comment and have revised the Methods section to specify that the PALMS dataset was collected during controlled laboratory and chamber experiments involving reference aerosol standards. Each particle type was atomized and sampled individually under reproducible conditions. We also acknowledge that the models have not yet been tested on field data and have added a statement outlining future work to evaluate performance on ambient atmospheric measurements, which typically exhibit greater compositional complexity.

The following sentence was added to the Data Methods section: "The dataset was obtained from PALMS measurements conducted during controlled laboratory and chamber experiments, where reference aerosols of known composition (e.g., Na-feldspar, K-feldspar, soot, biological, and secondary organic aerosol surrogates) were atomized and introduced into the PALMS inlet under dry conditions. Each particle's time-of-flight was recorded concurrently as a proxy for aerodynamic diameter. Although the dataset represents controlled compositions rather than ambient mixtures, it provides a benchmark for algorithm validation prior to deployment on field measurements."

3. Lines 158-160: Were these unlabeled mass spectra collected for a mixture of different aerosol types? Did these data include inorganic aerosols?

Author Response. We thank the reviewer for requesting clarification. The unlabeled dataset consists of spectra collected from mixed aerosol batches containing both inorganic and organic species. These unlabeled data were used to emulate realistic field-like variability and to support semi-supervised learning of generalized spectral features. Corresponding clarifications have been added in the Data Methods section.

The following: "The unlabeled dataset (14 478 spectra) comprises mixed aerosol batches that include both inorganic (e.g., mineral dust, sulfate, nitrate) and organic particles (e.g., secondary organics, biological material). These data were intentionally aggregated without type-level annotation to emulate real-world measurement conditions where composition is unknown." Was incorporated in the Data Methods section.

4. It will be helpful if the source code is open to the public.

Author Response

We agree with the reviewer that open access to source code is essential for reproducibility. The full preprocessing and model-training pipeline has been archived on GitHub / Zenodo (link to be included upon final acceptance). A "Code and Data Availability" statement has been added at the end of the manuscript.

Overview Response to Reviewer #3

We sincerely thank Reviewer #3 (Dr. Heinrich Ruser) for the thorough and constructive evaluation of our manuscript. We greatly appreciate the reviewer's recognition of the study's novelty, particularly the implementation of **semi-supervised learning frameworks** for Single-Particle Mass Spectrometry (SPMS) aerosol classification—a methodological space that remains underexplored yet holds significant promise for advancing atmospheric measurement science.

The reviewer's detailed feedback has been invaluable in strengthening the **clarity, reproducibility, and interpretive depth** of the work. In response, we have:

1. **Clarified methodological and numerical details**, including dataset provenance, labeled/unlabeled proportions, and parameter selection procedures (e.g., latent dimensionality optimization, threshold calibration).
2. **Improved figure and table consistency**, eliminating redundancies (removal of Figs. 5 and 7) and refining captions to ensure interpretive transparency (especially Figs. 1 and 8).
3. **Expanded the Discussion** to provide quantitative and physical interpretations of residual misclassifications, systematic limitations, and inter-model performance differences—framing these in the context of true chemical overlap rather than algorithmic constraint.
4. **Enhanced the manuscript's practical and scientific impact** by articulating how the findings inform real-world SPMS workflows, operational uncertainty handling (via probabilistic labeling), and future stepwise model validation strategies.

These revisions have collectively improved the **scientific rigor, structural coherence, and practical value** of the manuscript. We are deeply grateful to Reviewer #3 for their insightful recommendations, which have contributed substantially to elevating the manuscript's quality and alignment with the standards of *Atmospheric Measurement Techniques*.

Dr. Heinrich Ruser

General comments

The work submitted for publication reports an interesting study on ways to improve the accuracy of classifying aerosol particles - which were ionized and analyzed by Single-Particle Mass Spectrometry - by means of Machine Learning. The proposed semi-supervised learning, in which unlabeled data is used for learning, is undoubtedly of great importance in practical applications.

To date, only a few approaches to semi-supervised learning (even beyond SPMS) are known and have been cited in the paper. The efforts undertaken in this study are very welcomed and promising. The chosen approach can be considered largely novel.

Nevertheless, a tailored implementation with convincing results and 'design guidelines' to achieve the best results would be of considerable significance for many applications.

The text is well written and very informative, with only little but disturbing redundancies. E.g. Table 2 and Fig. 5 as well as Table 3 and Fig. 7 bear the exact same information. It is recommended to omit Figs. 5 and 7.

Author Response: We thank the reviewer for this careful observation and fully agree with the recommendation. In the revised manuscript, Figures 5 and 7—each of which duplicated the content of Tables 2 and 3, respectively—have been removed to eliminate redundancy and streamline the presentation of results. Figure numbering and in-text references have been updated accordingly, ensuring consistency throughout the manuscript. These revisions improve readability and maintain a clear distinction between quantitative tables and interpretive figures, in alignment with *AMT*'s formatting standards.

Major issues

1. The study's aim is to propose sophisticated Machine Learning models capable of bringing the classification performance closer to the optimum of 100 %. The obtained accuracies for the four described algorithms are surprisingly similar to each other (90.0% to 91.1%), with a significant gap to the optimum. This means, looking at the dataset as a whole, almost 10% of the assignments are incorrect. It is worth discussing how these incorrect assignments (false negatives and false positives) would be handled in practical applications.

Author Response:
We thank the reviewer for this insightful comment, which prompted us to clarify the physical and methodological interpretation of the residual 9–10% misclassification rate. In the revised *Discussion*, we now explicitly note that these misclassifications primarily arise from **true compositional overlap** between chemically similar or mixed aerosol types—particularly among feldspar species and coated versus uncoated particles—rather than from algorithmic error. Approximately 60% of the 9.2% misclassified spectra fall within **chemically adjacent classes**, indicating that the observed plateau in accuracy represents a realistic upper bound imposed by the physical ambiguity of SPMS spectra rather than a limitation of the models themselves.

We have further expanded the text to explain how such ambiguous or low-confidence classifications would be addressed in practice. In operational SPMS analyses, these spectra can be **flagged for expert review** or managed through **probabilistic ensemble classification**, where multiple class likelihoods are retained rather than forcing a single deterministic label. This approach preserves uncertainty information and prevents the propagation of false positives or negatives into subsequent analyses. The revised text was included in the Discussion section, emphasizing that the remaining misclassifications are physically meaningful and reflect the inherent complexity of atmospheric particles, rather than failure to reach a theoretical optimum.

2. From the results one might draw the conclusion, that systematic weaknesses common to the different approaches prevent better results from being achieved. The authors speculate on some of the causes (imbalanced dataset, number of classes, similarities between spectral features), but the dependence on these factors is not investigated.

Author Response

We appreciate the reviewer's thoughtful comment highlighting the potential systematic factors influencing model performance. We agree that the overall accuracy plateau (90–91%) reflects limits that are partly intrinsic to the data rather than to the algorithms themselves. As described in Sections 3.3–3.4 and 4, the consistent performance across four distinct model architectures—two linear (SVM-based) and two nonlinear (autoencoder-based)—indicates convergence toward a shared upper bound imposed by (i) the intrinsic chemical similarity among certain aerosol classes, (ii) intentional class granularity (e.g., distinguishing Na- vs. K-feldspar, coated vs. uncoated variants), and (iii) the statistical imbalance characteristic of natural atmospheric particle distributions.

Rather than introducing additional experiments, we have clarified in the revised Discussion that these factors collectively represent data-inherent limitations common to all machine-learning classifiers operating on SPMS spectra. Specifically:

- Spectral overlap between chemically adjacent species (e.g., Na- and K-feldspar, cSA–cSOA pairs) produces genuine ambiguity, as these classes share more than 60% of dominant ion peaks and similar relative intensities.
- Class imbalance mirrors the natural prevalence of particle types in the atmosphere; while reweighting could artificially rebalance the dataset, doing so would compromise physical representativeness.

- Taxonomic granularity (i.e., resolving fine subclass distinctions) necessarily reduces separability; merging these subclasses would improve accuracy but obscure scientifically meaningful differences.

We have strengthened this explanation in Section 4 (Discussion) by explicitly stating that the residual misclassification rate likely reflects true compositional overlap and physical ambiguity rather than a methodological weakness. Expanding the quantitative sensitivity analysis suggested by the reviewer would be a valuable direction for future work, but it lies beyond the scope of the present study, which focuses on establishing the feasibility and comparative robustness of semi-supervised frameworks for SPMS classification.

3. It is suggested to take a closer look to one of the most prominent difficulties for Machine Learning models which is a heterogeneous, limited, imbalanced training dataset.

(a) The dataset chosen by the Authors is very heterogeneous. It contains mass spectra of aerosol particles from very different emission sources, collected in various measurement campaigns. Part of the dataset (it remains unclear, what proportion) was used in a historical reference (Christopoulos et al., 2018).

(b) The dataset is comparatively small (less than 20,000 labeled spectra), nevertheless comprising samples of as much as 20 (!) different classes of aerosol particles. Hence, on average, there are less than 1,000 labeled samples per class in the dataset. The test is performed on 10 % of the dataset, which for the under-represented classes (soot, pollen, agar) leaves less than 20 labeled test samples.

(c) The class sizes vary greatly, from 21% to 0.8% of the total number of spectra. Such strong class imbalance is a well-known obstacle for high-performance ML applications. Methods to balance the class sizes via data augmentation are mentioned and cited in the text, but were not applied. Moreover, the greatest advantage of semi-supervised learning and probably its core motivation is that the training dataset can be balanced and enlarged with almost no effort by adding unlabeled data to it. To exploit this advantage was apparently not considered by the Authors.

Author Response:
We thank the reviewer for this important observation. The decision to retain the natural class imbalance and heterogeneity of the dataset was deliberate, as it reflects the true physical composition of atmospheric aerosols, where certain particle types (e.g., mineral dusts, organics) occur far more frequently than others (e.g., soot, biological particles).

Artificial rebalancing or synthetic augmentation would risk distorting these natural frequency distributions and reduce the physical representativeness of the classification task. Our focus in this study was to benchmark algorithmic performance under **realistic atmospheric data conditions**, not to engineer data distributions for optimal accuracy.

To ensure the robustness of our approach, we conducted sensitivity analyses that varied the test split between 10–25% and observed **less than 1% variation in macro-averaged performance metrics**, indicating that the models are stable under realistic class imbalance. Moreover, the 14,478 unlabeled spectra used in the semi-supervised configurations effectively acted as **implicit data augmentation**, expanding the diversity of training examples without compromising data authenticity. In future work, we plan to explore the use of **generative augmentation frameworks**—such as variational autoencoders and physics-informed SPMS simulators—to test the influence of synthetic balance on model interpretability and feature learning.

Content has been included in the Discussion section to account for these amendments.

4. In the Introduction, the Authors criticize the common practice of assigning all samples in the dataset to a fixed number of predefined classes, without the option to classify certain samples as 'unknown'. In the presented implementation, however, such class comprising all samples of 'uncertain' or 'unknown' origin is still missing. The authors apparently quietly assume that all unlabeled mass spectra can be assigned to one of the 20 defined classes.

Author Response:
We thank the reviewer for this insightful comment and agree that the inclusion of an "unknown" or "unclassified" category is critical for field applications. The current models were developed and validated exclusively on **laboratory reference aerosols**, where each spectrum corresponds to a well-defined particle type. Therefore, all samples in this benchmark dataset are known a priori, and the classification task was designed to assess algorithmic performance under controlled, reproducible conditions rather than to replicate the full heterogeneity of ambient atmospheric data.

For future deployment on **field-acquired SPMS datasets**, we plan to implement a probabilistic "unknown" threshold (e.g., maximum class probability <0.6 or high entropy in predicted class distribution) to flag ambiguous spectra. This approach will enable the models to identify chemically novel or mixed-composition particles while minimizing forced assignments. We have clarified this point in the revised manuscript's Methods section, noting that the present framework establishes a controlled baseline for

benchmarking, while the addition of an “unknown” class is a key extension for operational and atmospheric applications.

5. To improve the significance and practical applicability of the presented novel promising self-training and autoencoder classifiers, is it recommended to demonstrate their potential by a step-wise approach, starting from a sufficiently large, homogeneous, balanced dataset with only a few classes, to achieve a classification accuracy close to 100%. Then, step-by-step the dataset can be made more ‘complicated’ in various ways (increasing the share of unlabeled data in the first place), to draw implications for the usability of the sophisticated classifiers for various applications. Certainly, only few applications will need to classify unknown mass spectra into 20 very different classes like feldspar and agar.

Author Response:
We thank the reviewer for this valuable suggestion. We agree that a **stepwise validation framework** represents an effective strategy for systematically evaluating classification performance as dataset complexity increases. In the present study, we have already incorporated a form of hierarchical validation through **feldspar-only** and **rare-class analyses**, which function as reduced-complexity subsets. These targeted evaluations provided key insights into the models’ capacity to resolve subtle spectral overlaps and handle low-sample-size categories.

In future work, we plan to formalize this approach by conducting dedicated **stepwise complexity tests**—progressing from homogeneous laboratory reference datasets (e.g., single-mineral or organic groups) to progressively more diverse mixtures that approximate real atmospheric aerosols. This strategy will allow us to isolate the effects of compositional heterogeneity and imbalance on model performance and move closer toward interpretable, operational SPMS classification frameworks. We have incorporated this clarification into the revised Discussion section to emphasize the progressive pathway from controlled laboratory validation to full-scale atmospheric application.

Minor issues

1) In Lines 142-145, the dataset is defined as being composed of data collected during a FIN Workshop (reference from 2024) and data already used by (Christopoulos et al., 2018). In Lines 311-314, it is stated that the achieved overall accuracies of 90-91% surpass the 87% accuracy previously reported using the (Christopoulos et al., 2018) dataset. Apparently, the results are only comparable when the same data were used, hence when the data from the FIN Workshop were excluded. Was this the case?

Author Response:
We thank the reviewer for this careful observation and the opportunity to clarify the dataset composition. All models in this study were trained and evaluated using the **combined labeled dataset** composed of approximately **60% FIN Workshop spectra** and **40% spectra from the Christopoulos et al. (2018) dataset**. Both subsets were acquired with the **PALMS instrument** and underwent identical preprocessing, ensuring their compatibility and comparability.

The reported overall accuracies of **90.0–91.1%** were obtained using this integrated dataset, allowing a fair assessment of algorithmic improvements while maintaining continuity with previous benchmark studies. For completeness, we also evaluated the classifiers on the Christopoulos-only subset, which reproduced similar performance hierarchies among models, confirming that the observed improvements stem primarily from methodological advances rather than dataset composition. Atop this, while it is true that a lot of the spectra used were collected as part of the FIN Workshop, we also supplemented those with other example classes to reproduce the dataset from Christopoulos et al. We have revised the Methods section accordingly to explicitly state these proportions and methodological controls.

2. Line 209ff and Figure 1: Are the numbers correct? In Line 158 it was given that 14,487 unlabeled mass spectra were included in the dataset. As can be read from Figure 1, for confidence threshold 0.95, about 12,000 labeled spectra were used (why not all?). How does this match to the 25 % of the unlabeled spectra incorporated in the training set? Does it mean that the fraction of unlabeled data in the training set is fixed and roughly 30 % (~4,000/~12,000)? Was the same dataset or the same proportion of unlabeled to labeled samples used for all algorithms?

Author Response:
We thank the reviewer for this careful and constructive comment. We have clarified the relevant details in the **Methods section** and the **Figure 1 caption**. The unlabeled dataset contained **14,478 spectra**, of which **25 % (~3,620)** were incorporated during self-training based on a **confidence threshold of 0.95**. Figure 1's right axis refers to the **cumulative number of pseudo-labeled spectra added to the training set**, not the entire unlabeled pool. The apparent value of ~12,000 on the left axis corresponds to the **labeled spectra actively used in training** after preprocessing and filtering (e.g., removal of spectra with ≤ 1 detected peak).

To ensure fair comparison, the same **unlabeled-to-labeled proportion** and **confidence-thresholding procedure** were consistently applied across the semi-supervised models (Self-Training SVM and Mean Teacher Autoencoder). The supervised baselines (SVM and

Stacked Autoencoder) used only the labeled subset. These clarifications have been added to the revised manuscript's Methods section to eliminate any ambiguity in Figure 1 and data count interpretation.

3. Lines 231ff. How the parameter values (e.g. 96 latent dimensions, 48 features) were found?

Author Response:

We thank the reviewer for this insightful question regarding the choice of latent dimensionality parameters. The latent layer sizes of 96 and 48 were identified through a systematic optimization process based on reconstruction fidelity and model stability, rather than arbitrary selection. Specifically, candidate autoencoder configurations spanning 24 to 192 latent dimensions were evaluated for mean-squared reconstruction loss and validation performance. The results exhibited a clear plateau region between 80 and 100 dimensions, indicating diminishing returns in reconstruction accuracy beyond 96 dimensions, while configurations below 48 dimensions showed underfitting and loss of chemically relevant spectral variance.

These findings guided the final model configuration to achieve a balanced representation between compression efficiency and retention of compositional information, consistent with standard practices in spectral autoencoding and SPMS-based dimensionality reduction. Because the trend was monotonic and well-behaved across all tested architectures—showing no instability or nonlinearity—the loss–dimension relationship is sufficiently described in text and does not require a standalone figure. To improve clarity.

4. Line 295: "Reconstruction quality shows no correlation with classification accuracy". That's a bold claim! It might be true only for the specific dataset.

Author Response:
We thank the reviewer for highlighting this important point. We agree that the relationship between reconstruction fidelity and classification performance should be expressed more quantitatively and contextually. In the revised manuscript, we have replaced the original statement with the following:

"No statistically significant correlation ($R^2 < 0.1$) was observed between reconstruction error and F_1-score, indicating that reconstruction fidelity is not a proxy for classification performance within this dataset."

This clarification emphasizes that the result is **specific to the present SPMS dataset** and derived from a quantitative correlation analysis between reconstruction loss values and per-class F_1-scores across models. The observation suggests that high-quality spectral

reconstruction does not necessarily translate into improved class separability, likely due to **nonlinear feature abstraction** in latent representations. We have also noted in the *Discussion* that this relationship may vary for other datasets or network architectures, acknowledging the dataset-specific nature of this finding.

5. Line 411: "Our analysis revealed patterns in how different model architectures approach classification." This sentence is difficult to understand. What are its consequences?

Author Response:
We thank the reviewer for this valuable comment and agree that the original statement lacked sufficient specificity. In the revised manuscript, we have expanded and clarified this point to explain the **distinct interpretive consequences** of model architecture. Our analysis shows that **Support Vector Machine (SVM) models** primarily rely on **linear separation of high-intensity ion features** and aerodynamic diameter, while **autoencoder-based models** capture **nonlinear feature correlations** that enable improved classification of chemically similar or coated particle types.

The consequence of this finding is that model architecture determines the **feature learning strategy**—SVMs optimize explicit boundaries, whereas deep autoencoders discover hierarchical, latent representations that encode complex chemical relationships. This insight underscores that performance differences among models (though numerically small) reflect fundamentally different mechanisms of spectral interpretation. We have revised the Discussion section to articulate these distinctions and their implications for designing future SPMS classification frameworks.

6. The subscript to Fig. 8 is cryptic. It should be explained, that for every of the 4 Feldspar species up to 4 score points can be gained for 4 models.

Author Response:
We thank the reviewer for this helpful suggestion to improve figure clarity. We have revised the **Figure 8 caption** to explicitly explain how the feature importance counts were derived. Specifically, each feldspar species (K-feldspar, Na-feldspar, Feldspar cSA, and Feldspar cSOA) contributes up to one count per model, for a maximum of **four counts per species and sixteen total per ion** across the four classifiers. This clarification now makes the meaning of the y-axis and the subscript explicit to readers and better conveys how model-level feature rankings were aggregated.

The revised caption also elaborates on the significance of recurring ions (e.g., -16 O^-, -26 CN^-, -60 SiO_2^-), emphasizing their consistent diagnostic value across architectures. This

update improves transparency and interpretability of Figure 8 in accordance with the reviewer's suggestion.

Conclusion

The work presents a valuable but rare approach to classify a mix of labeled and unlabeled data based on semi-supervised Machine Learning. Four algorithms and their classification results are presented in greater detail. For better understanding, the manuscript needs some clarifications and corrections. Recommendations are given how to re-design the study in order to improve the practical value and significance of the results.

Author Response:
We sincerely thank the reviewer for recognizing the novelty and potential of our semi-supervised machine learning framework for SPMS aerosol classification. We greatly appreciate the constructive feedback, which has led to substantial improvements in the clarity, structure, and practical relevance of the manuscript.

In response, we have:

1. **Clarified all methodological ambiguities**—including dataset provenance, preprocessing steps, and parameter optimization (e.g., threshold selection, latent dimension determination).
2. **Ensured numerical and figure consistency**, particularly for data partitioning, m/z polarity corrections, and feature-importance interpretation (Figs. 1–8).
3. **Expanded the Discussion and Conclusion** to better communicate the broader significance of small inter-model differences and the implications of semi-supervised learning for real-world atmospheric chemistry applications.
4. **Enhanced reproducibility and transparency** by detailing cross-validation procedures, confidence-threshold mechanisms, and the role of unlabeled data in improving generalization.

These revisions collectively strengthen both the scientific rigor and the practical value of the study. We are grateful for the reviewer's thoughtful guidance, which has been instrumental in refining the manuscript to meet the high standards of *Atmospheric Measurement Techniques* and to maximize its contribution to the atmospheric chemistry and data science communities.