**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 (R² < 0.1) was observed between reconstruction error and F₁-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 F₁-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 SiO₂⁻), 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.