General Comments:

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.

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.

Specific Comments:

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.

2. Line 81

Add and cite recent research works.

3. Line 83

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

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.

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.

2. 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. 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?

3. 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?

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.

4. 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.

5. 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?

6. 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.

7. Line 182

Citation error, remove (Christopoulos et al., 2018)

Section: Results

1. 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. 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.

2. Line 340

How did you analyze the importance of the ions?

3. Line 346 - 358

Error analysis need some mass spectra as example.

4. Line 374, Line 411

Have you trained models separately with and without aerodynamic diameter and compared their results? The aerodynamic diameter accounts for only one feature out of 194, so its relative weight is just 1/194.