Supplementary Information

for

Max-Intensity Untargeted Transformation (MINUT) for Direct Chemometric Modeling of High-Resolution Mass Spectrometry Data

DBPR

Baseline

To compare our method, we implemented several baseline and alternative binning strategies. All comparative models were processed under identical normalization and scaling procedures to maintain fair evaluation conditions, using the same model parameters and the same number of selected features. We applied min–max intra-spectrum normalization for intensity (I), and trained a random forest classifier with 100 estimators on 100 selected features. These comparative methods aimed to evaluate the impact of using the maximum versus the average intensities in the binning process and to compare to classic untargeted method. This design enabled us to assess whether selecting the maximum intensity, rather than averaging intensities, influenced classification outcomes, and whether integrating both m/z and retention time data enhanced performance, compared to a classical one-dimensional baseline.

1D classic baseline intensity with mean data binning To benchmark our classification framework against traditional preprocessing approaches, we implemented a standard one-dimensional binning strategy based solely on the m/z dimension. This approach ignores retention time and constructs features by aggregating signal intensities within discrete m/z intervals for each sample. Specifically, for each sample, all spectra were pooled by concatenating m/z and intensity values across scans regardless of retention time. The m/z range from 100 to 1000 was divided into 900 uniform bins of 1 m/z width. The selected bin size ensures a comparable data reduction to that achieved by MINUT on this dataset. Each detected m/z value was assigned to its corresponding bin, and the mean intensity was calculated per bin. This produced a fixed-length intensity vector representing each sample, which served as input for downstream classification models. This method corresponds to a commonly used baseline in untargeted LC-HRMS classification task, where spectral structure over retention time is not explicitly modeled.

1D classic baseline intensity with max data binning To evaluate the effect of using maximum intensity rather than mean intensity for each bin, we applied the same one-dimensional binning strategy as above, but replaced the mean aggregation with the maximum intensity per bin. This alternative allowed us to assess how different bin-level aggregation strategies (maximum vs. average) influenced classification performance. All comparative models used the same normalization and scaling procedures to ensure consistency.

2D binning with average intensity Finally, to compare our method against a two-dimensional binning approach, we adopted the same 2D strategy as in MINUT, but instead of retaining the maximum intensity along with its associated m/z and retention time, we computed the average

of all intensities within each two-dimensional bin. This allowed us to evaluate whether MINUT outperforms classical methods not only due to the use of 2D binning, but also because the max-binning strategy enables preservation of both m/z and RT information.

Guidelines for Applying MINUT

One of MINUT's major strengths is its technique-independence: the framework operates directly on the raw mass–retention time–intensity space without requiring assumptions about ionization modes, matrix type, or detector type. As such, MINUT can be applied to a wide variety of analytical techniques. However, different chromatographic–HRMS setups can have different typical retention time (RT) variabilities. The following recommendations are intended to help practitioners by providing practical guidelines covering bin size selection, retention time considerations, feature retention, and validation strategies.

Bin Size Selection

- **m/z axis:** Depending on data complexity, the optimal m/z bin size will vary. Begin with a larger bin size to identify a parsimonious solution; if necessary, reduce the bin size to maximize the number of distinct compounds.
- **RT axis:** Choose bins at least three times the expected retention time variability for a given compound (based on the technique/column) to ensure an entire peak remains within a single bin across samples.

Retention Time Considerations

- No RT correction is required if category–batch distribution is balanced; otherwise, batch effects could bias results. In such cases, restrict the RT range or apply correction as needed, or alternatively rely solely on m/z.

Feature Retention

- Retain m/z and RT for category-specific biomarker discovery; include intensity only if quantitation or abundance trends are of interest.

Cross-Validation

- Select bin sizes and perform feature selection based on cross-validation to avoid overfitting and to ensure stability.

Interpretation

- Combine Gini score ranking with univariate statistics (e.g., Fisher's exact test) for robust biomarker selection.
- Once the most discriminative bins are identified, plot both multivariate and univariate analyses to detect category clusters and subsequently perform chemical interpretability based on the clustered m/z and RT values.

These recommendations are intended to ensure that MINUT delivers stable, fast, interpretable, and reproducible results across diverse chromatographic–HRMS configurations.

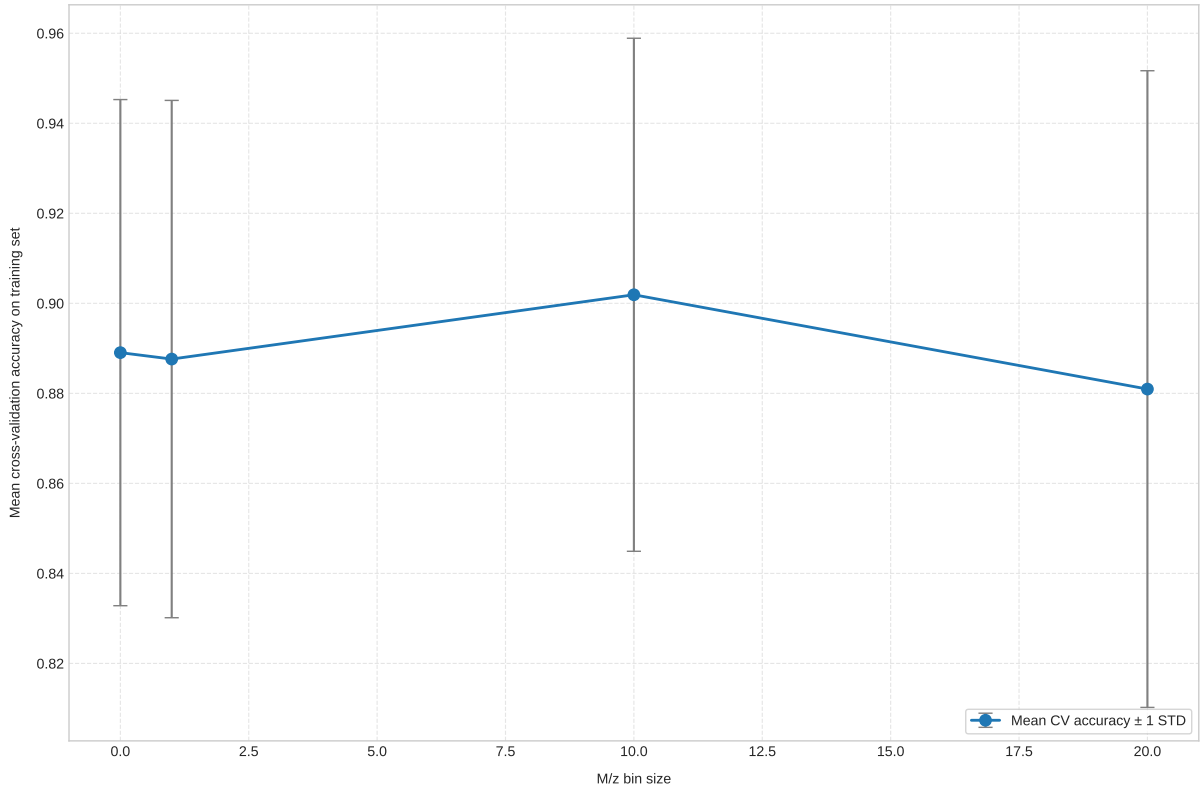

Figure S1: Influence of m/z bin size on mean cross-validation accuracy. Preliminary testing was conducted on the milk dataset's training set using the 100 most discriminative features.

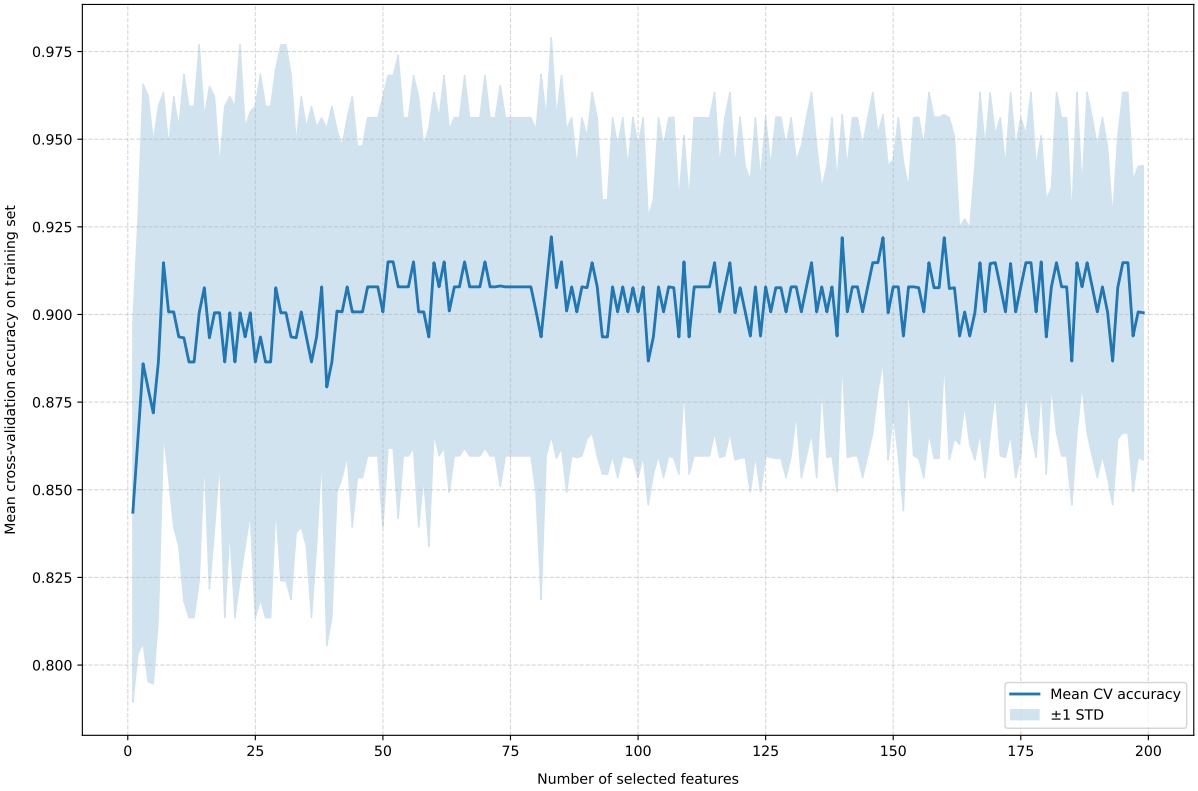

Figure S2: Influence of number of selected features on mean cross-validation accuracy. Preliminary testing was conducted on the milk dataset's training set.

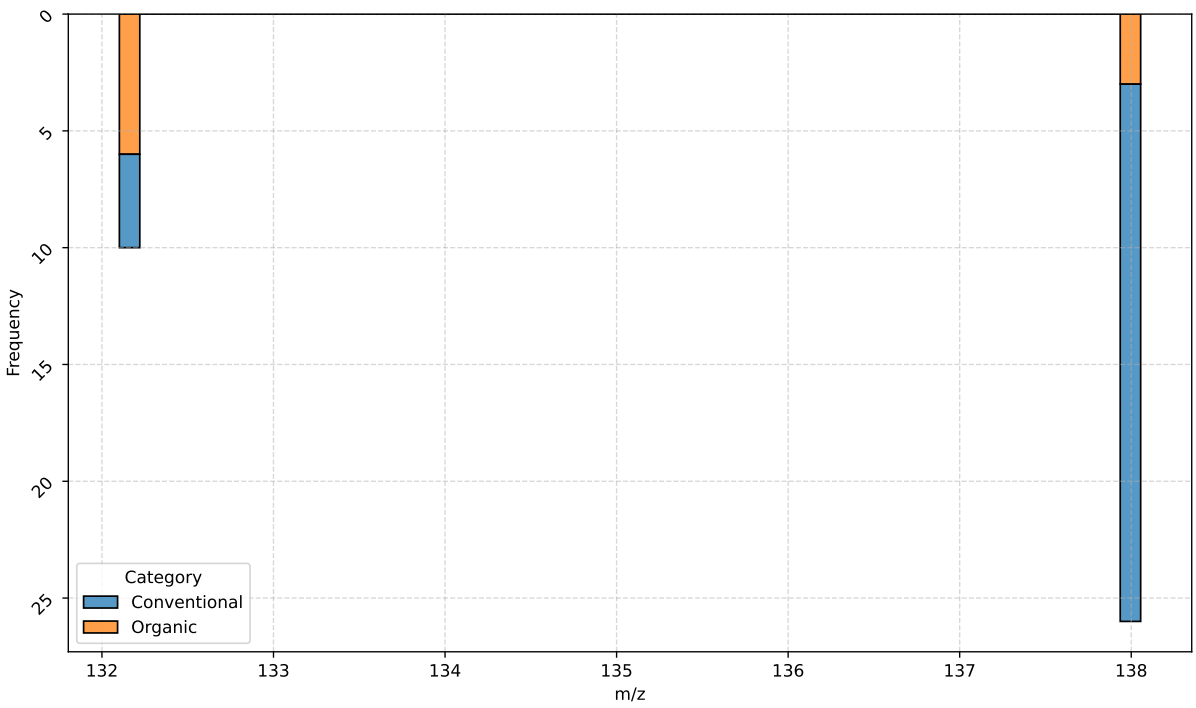

Figure S3: M/z histogram for the selected bin (defined as $m/z \in [130, 140)$ and retention time $\in [6, 8)$ minutes), colored by category : blue for conventional and orange for organic, based on the overall dataset.

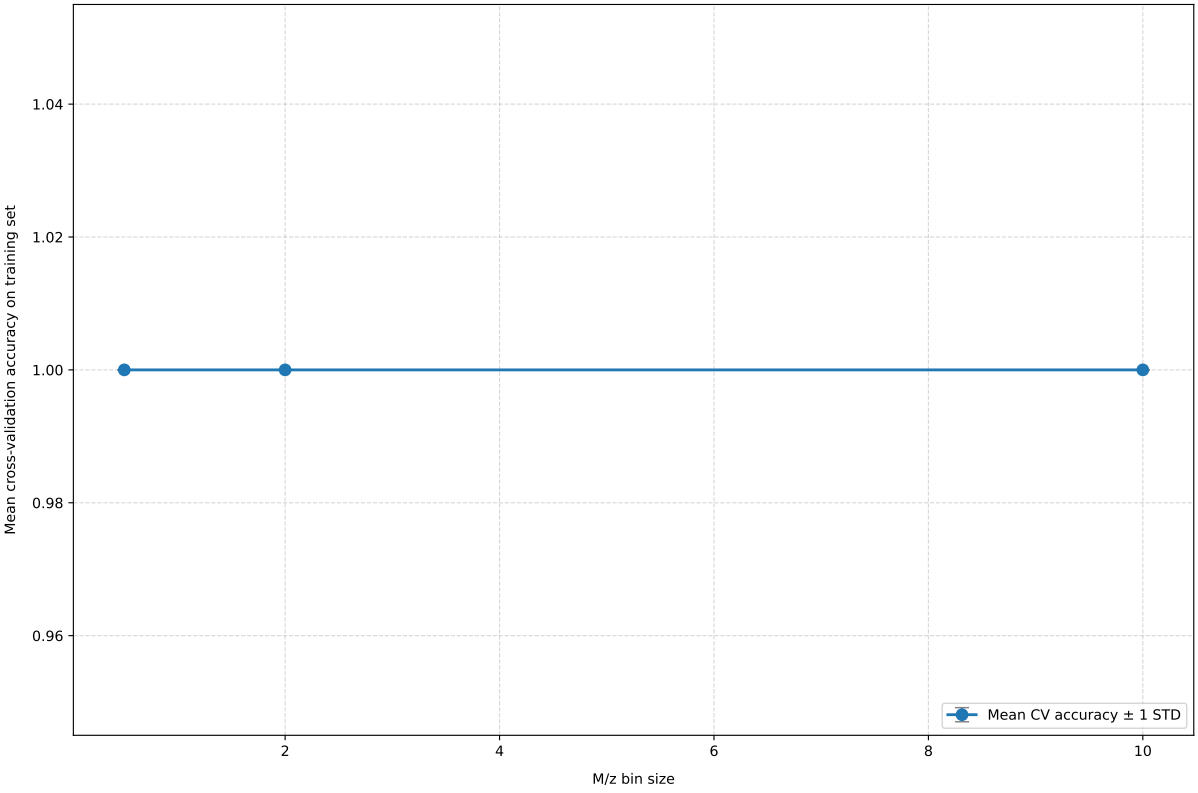

Figure S4: Influence of m/z bin size on mean cross-validation accuracy. Preliminary testing was conducted on the lung dataset's training set using the 50 most discriminative features.

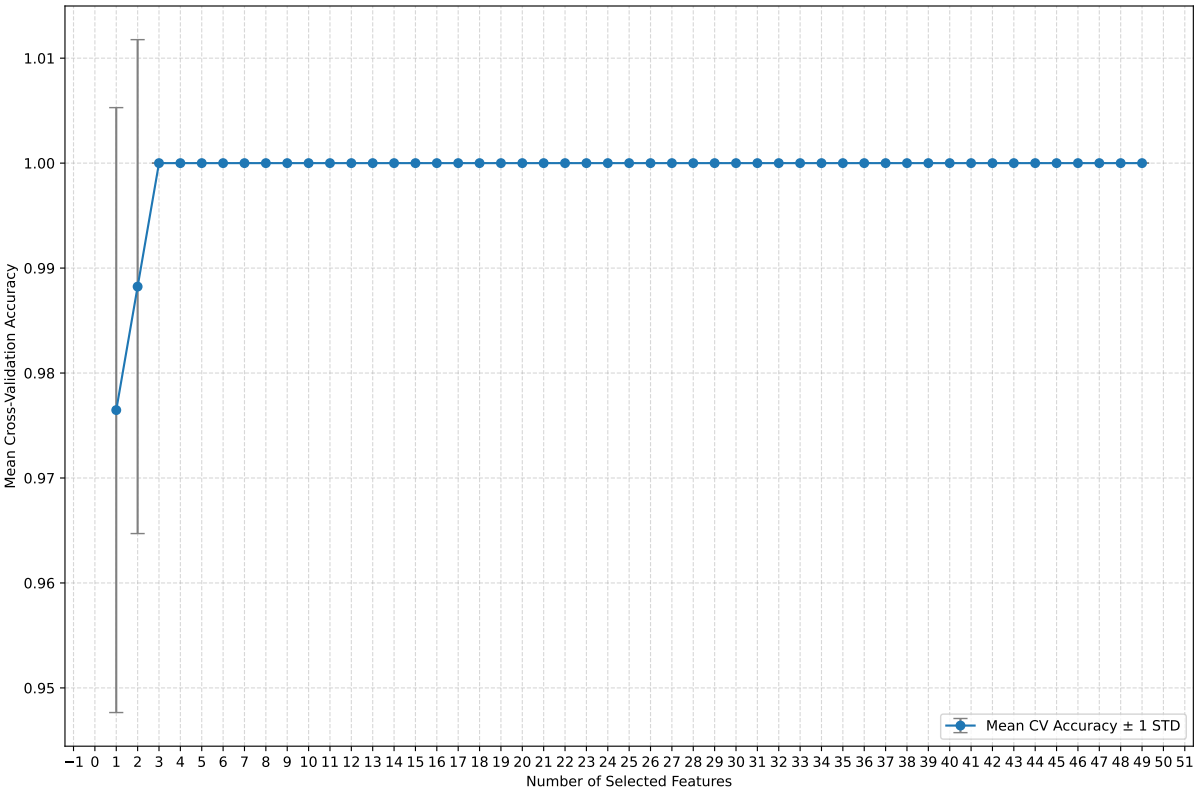

Figure S5: Influence of number of selected features on mean cross-validation accuracy. Preliminary testing was conducted on the lung dataset's training set.

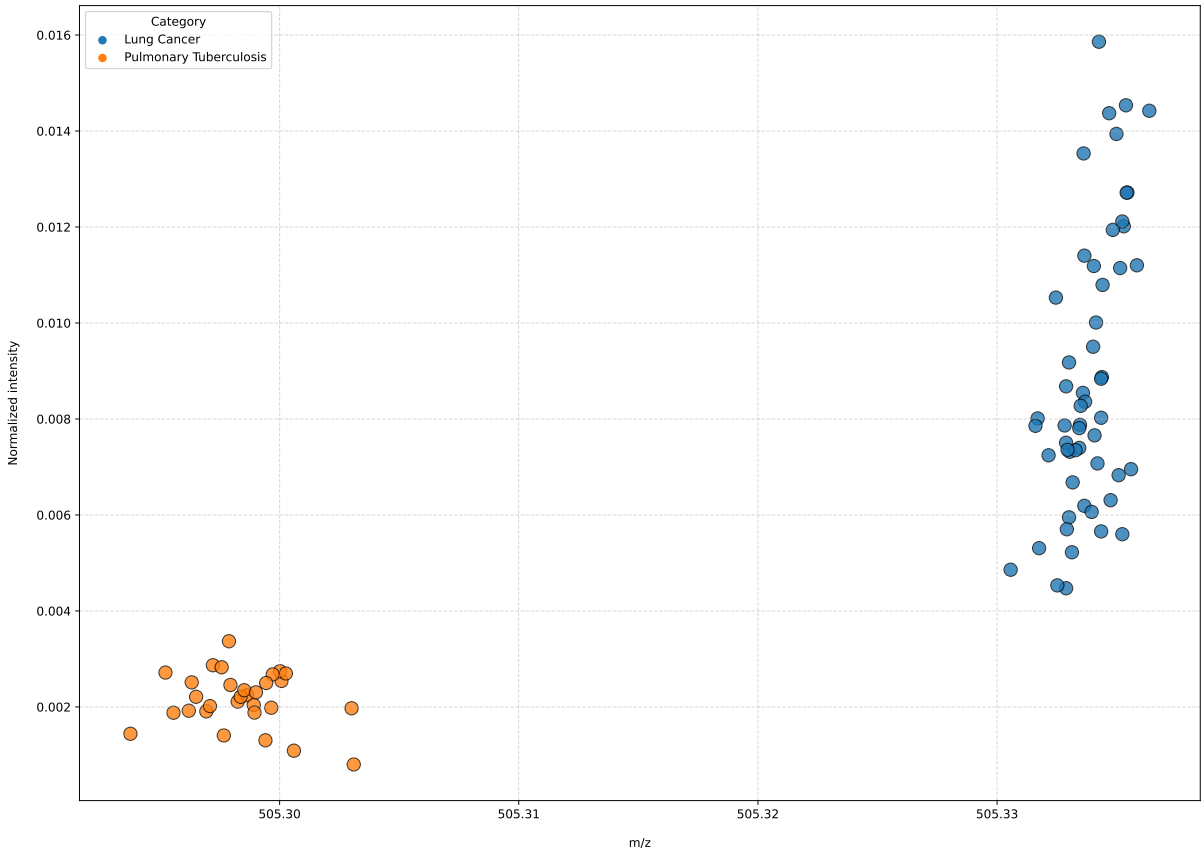

Figure S6: Relationship between normalised Intensity and m/z values (505.29–505.34 m/z) for the selected bin, colored by two forms of lung pathology. Based on the train dataset: blue represents lung cancer and orange represents pulmonary tuberculosis.

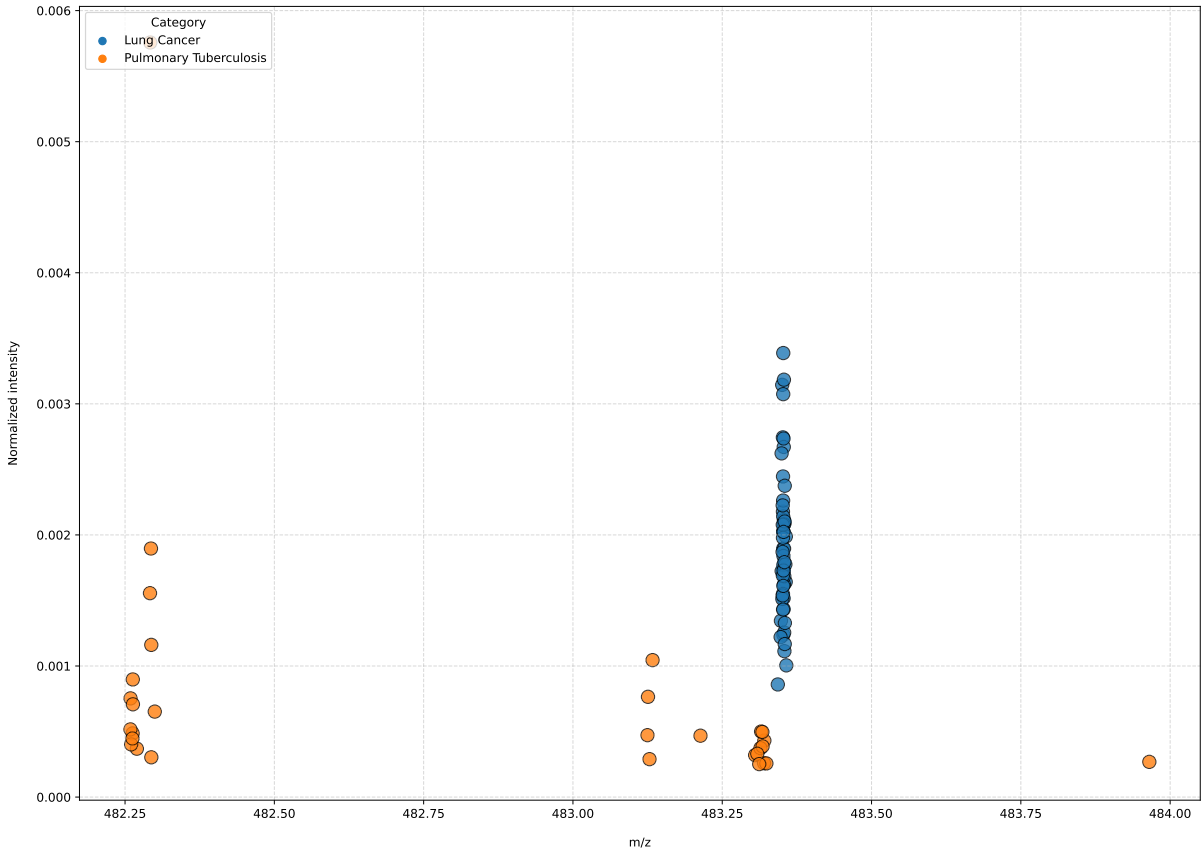

Figure S7: Relationship between normalised Intensity and m/z values (482.25–484 m/z) for the selected bin, colored by two forms of lung pathology. Based on the train dataset: blue represents lung cancer and orange represents pulmonary tuberculosis.

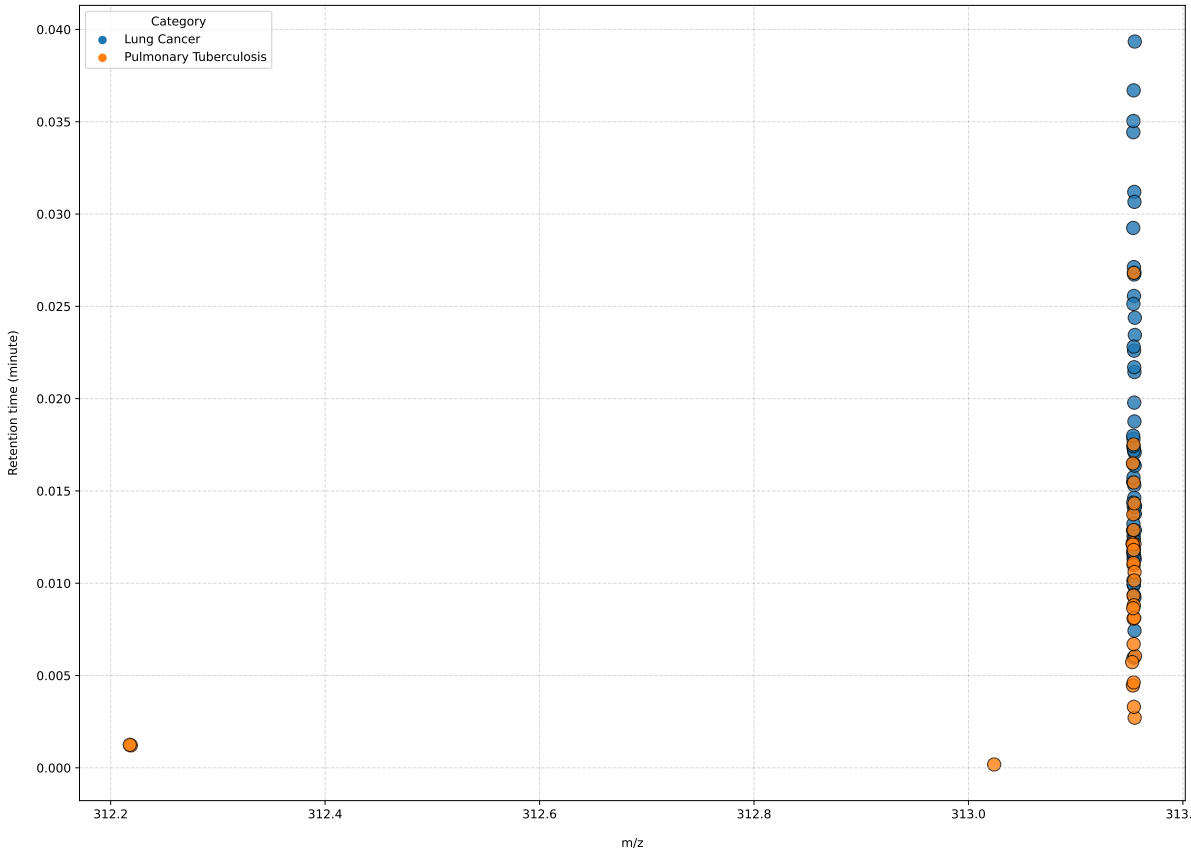

Figure S8: Relationship between normalised Intensity and m/z values (312.2–313.2 m/z) for the selected bin, colored by two forms of lung pathology. Based on the train dataset: blue represents lung cancer and orange represents pulmonary tuberculosis. The 313.1542 m/z cluster corresponds to phenylalanylphenylalanine.

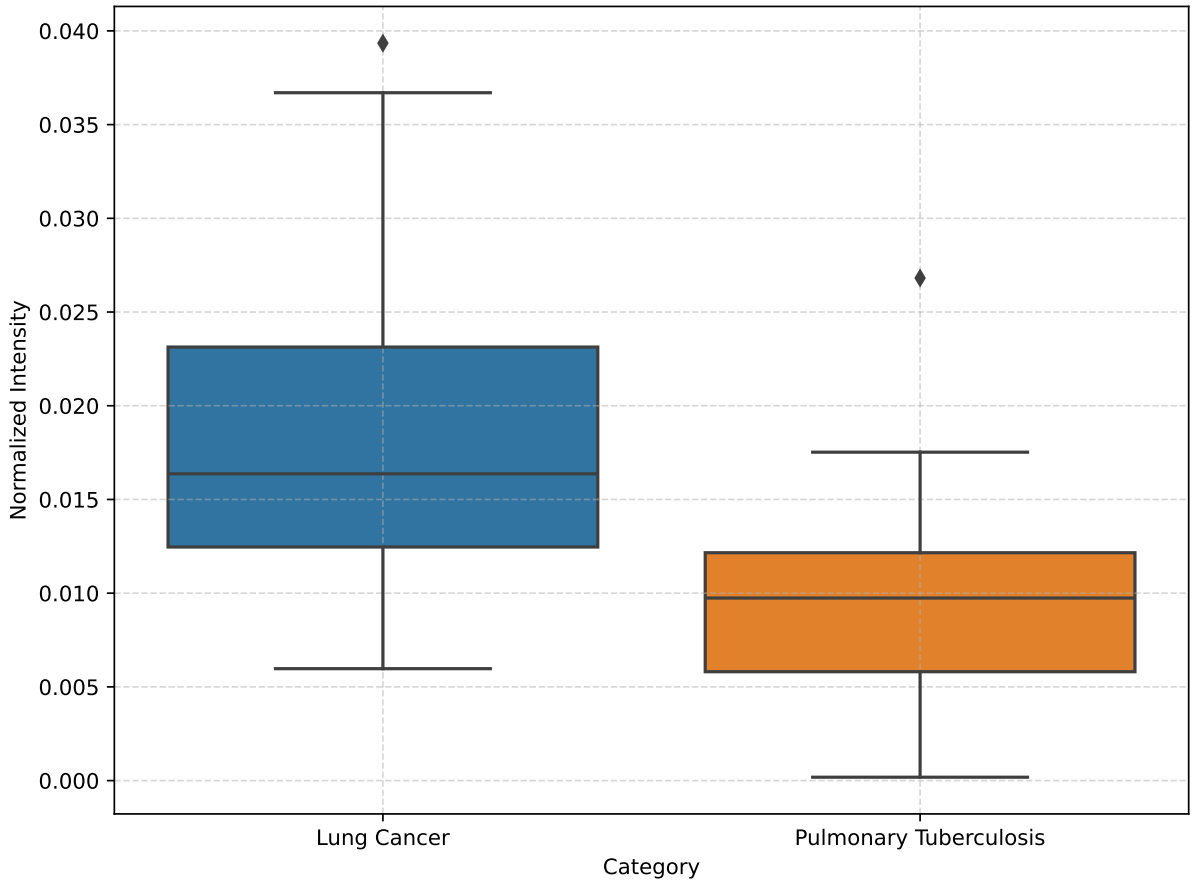

Figure S9: Boxplot of intensity for phenylalanylphenylalanine cluster (313.1542 m/z) on the discriminative bin, colored by category. Based on the test dataset: blue represents lung cancer and orange represents pulmonary tuberculosis.

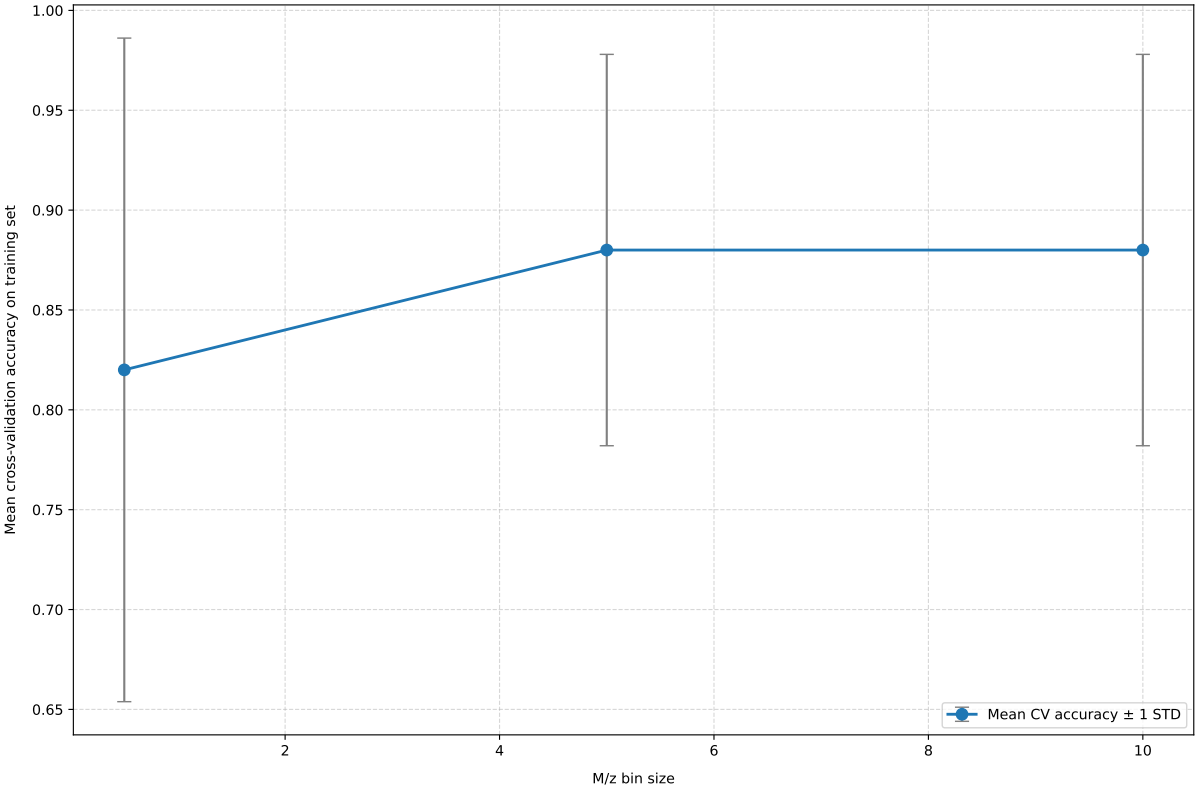

Figure S10: Influence of m/z bin size on mean cross-validation accuracy. Preliminary testing was conducted on the covid dataset's overall set using the 200 most discriminative features.

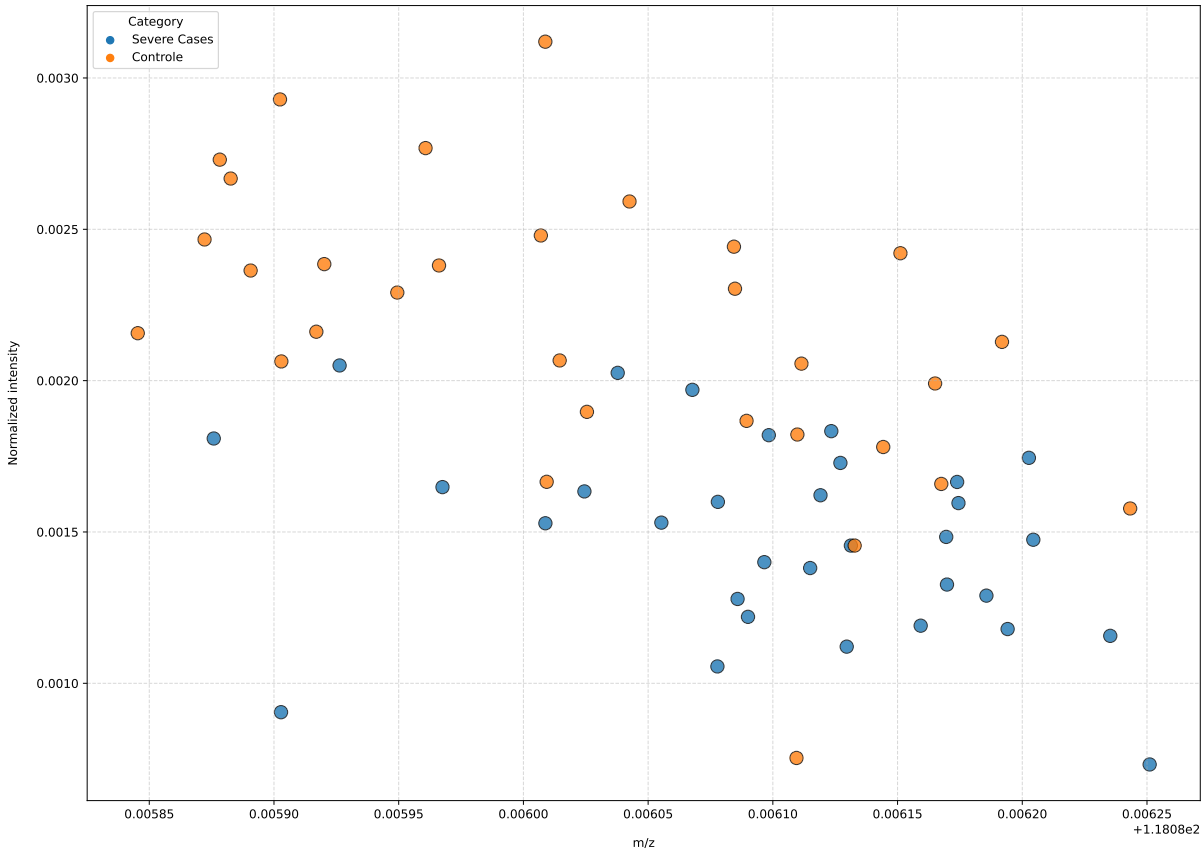

Figure S11: Relationship between nomalized intensity and m/z values (118.08580–118.08630 m/z) for the selected bin, colored by COVID-19 cases. Based on the overall dataset: blue represents severe and orange represents control cases.

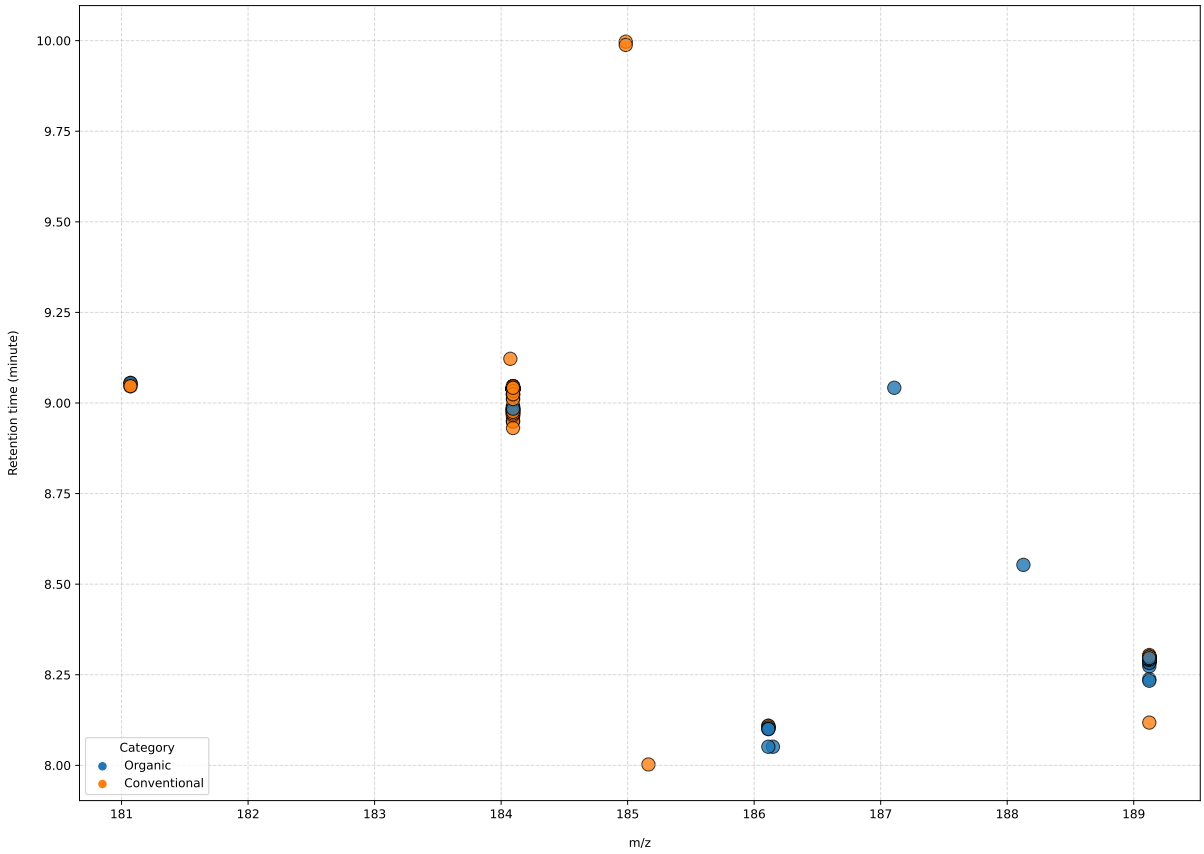

Figure S12: Relationship between retention time (8–10 min) and m/z values (181–190 m/z) for the selected bin, colored by milk category. Based on the overall dataset: blue represents conventional milk and orange represents organic milk.

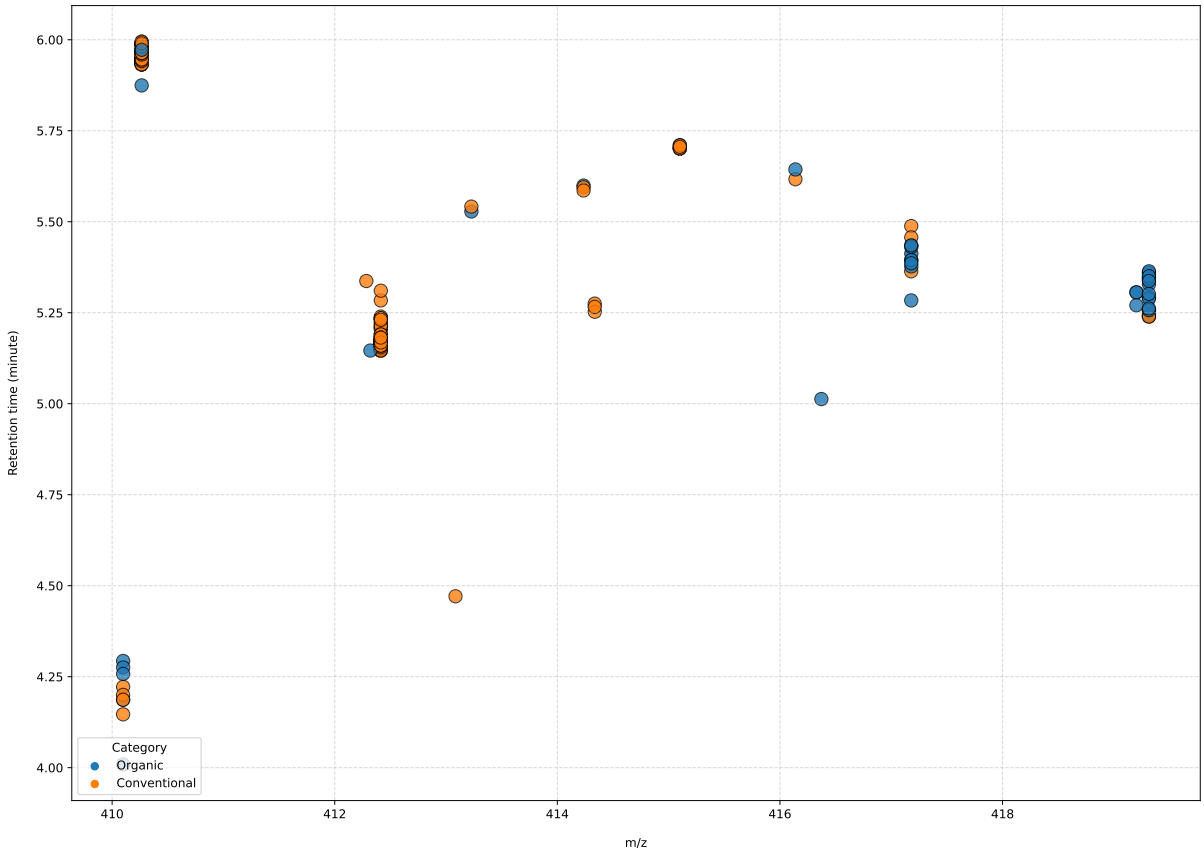

Figure S13: Relationship between retention time (4–6 min) and m/z values (410–420 m/z) for the selected bin, colored by milk category. Based on the overall dataset: blue represents conventional milk and orange represents organic milk.

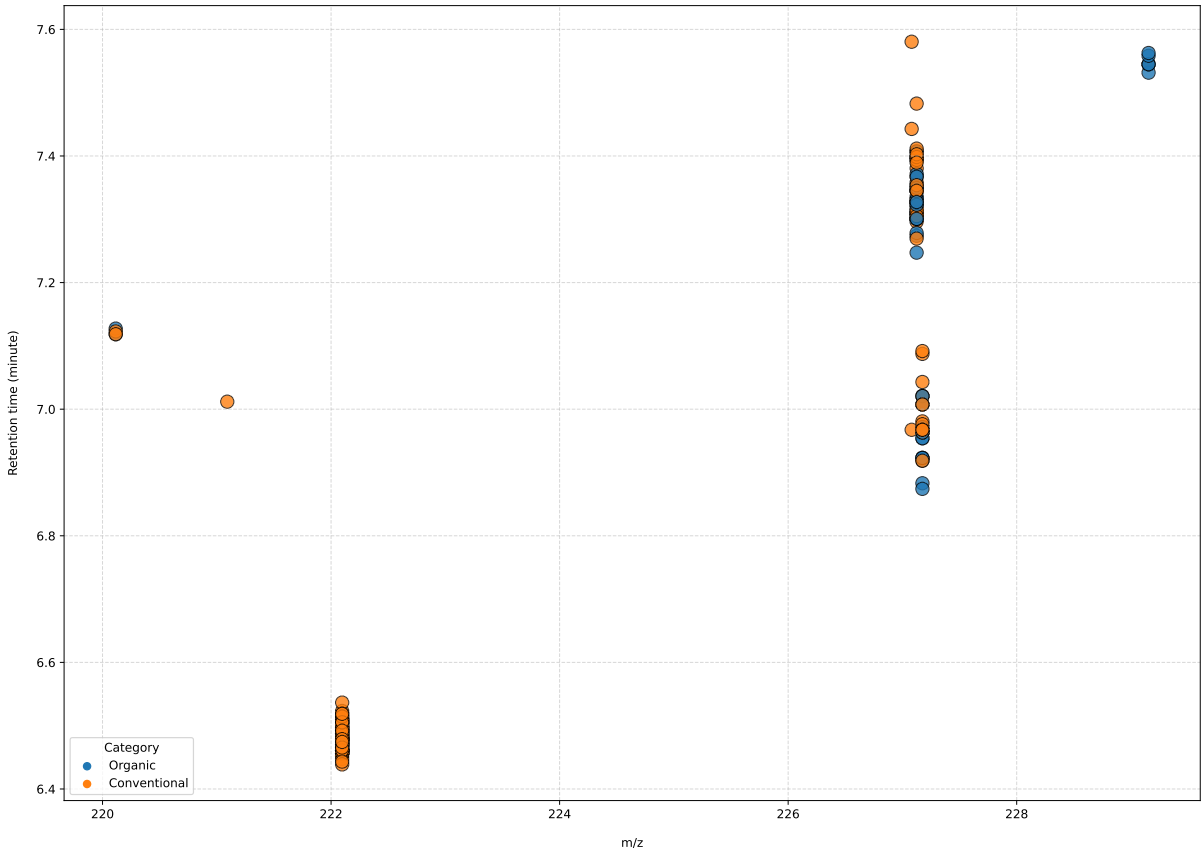

Figure S14: Relationship between retention time (6.4–7.6 min) and m/z values (220–230 m/z) for the selected bin, colored by milk category. Based on the overall dataset: blue represents conventional milk and orange represents organic milk.