I divided the text in three sections, following the data processing workflow:
1. Standardizing MS2 Spectra Formatting for Structured Data Processing.
2. Resolving MS2 Redundancy and Integrating MS1 Data for an MS2-Based Feature Table.
3. Align, Filter, and Refine: Identifying Features in Experimental Data

Please also find the detailed documentation for the whole workflow in the GitHub repository: https://github.com/EdwinChingate/feat.-ms2-Gauss/wiki/Functions-index, including a description for each function and variable in the code.

Standardizing MS2 Spectra Formatting for Structured Data Processing

A well-structured data framework ensures consistency, enabling meaningful statistical analysis, reproducibility, and reliable interpretation. The first step in most data processing workflows involves converting raw data into an open format, such as .mzML, which facilitates seamless data retrieval and analysis. .mzML is an open, XML-based format that has become a standard for mass spectrometry data, allowing compatibility across software tools (Deutsch, 2010). msconvert, a tool from the ProteoWizard software suite, converts proprietary .raw files from Thermo Scientific into .mzML, enabling integration with diverse software tools and programming languages, such as Python, for further processing (Adusumilli and Mallick, 2017). Once converted, LC-HRMS-derived data contain thousands of molecular features structured within multi-dimensional datasets (Renner et al., 2022). These datasets consist of spectra recorded at different elution times, capturing the dynamic composition of the sample over the chromatographic run. Each recorded spectrum contains millions of ion signals, varying in _m/z_ and intensity. The signals contained in MS1 and MS2 spectra are complemented by metadata such as precursor descriptors, which provide contextual information for further analysis. While MS1 spectra provide an overview of the abundance and _m/z_ for all detected ions, MS2 spectra capture the fragmentation patterns of selected ions, typically the most abundant in MS1 (Guo & Huan, 2020).

The Orbitrap mass analyzer captures ions in an electrostatic field and measures their oscillations as image currents over a defined period (Eliuk & Makarov, 2015). A Fourier Transform converts these time-domain signals, representing dynamic current intensities, into mass spectra: MS1 for all detected ions and MS2 for fragments produced via higher-energy collisional dissociation (HCD) (Eliuk & Makarov, 2015). In the Orbitrap, thousands of ions oscillate at frequencies inversely proportional to the square root of their m/z (Perry et al., 2008). As thousands of ions oscillate simultaneously, their collective perturbations average out, producing intensity variations that statistically conform to a Gaussian distribution centered around each ion's _m/z_ value (Figure 1.). This process results in distinct peaks in the mass spectrum, aligning with the principles of the central limit theorem. Each acquisition cycle yields a corresponding spectrum, reflecting the distribution of trapped ions as a series of peaks at characteristic _m/z_ values.

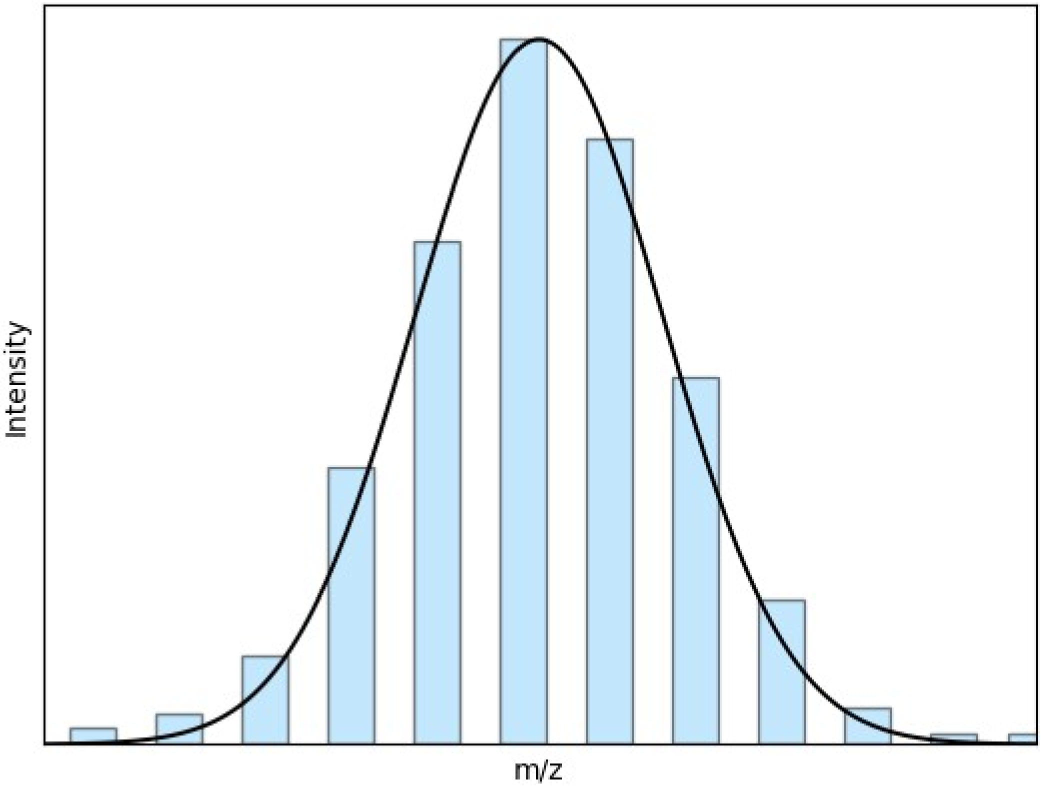

Figure 1. Describing a distribution of spectral signals with a Gaussian.

Gaussian modeling provides a structured approach to peak characterization, ensuring consistent peak representation and facilitating systematic comparisons across spectra (Renner & Reuschenbach, 2022). The ms2Gauss workflow identifies peaks as clusters of spectral signals following a normal distribution. Within this framework, the ms2Peak (MS2) and mzPeak (MS1) functions extract peaks, which represent ions and serve as the core analytical units of the dataset, facilitating both meaningful interpretation and streamlined downstream data processing. Furthermore, the ms2_spectrum function sequentially applies ms2Peak, integrating extracted peaks into a Spectrum table (Table 1.) to standardize peak representation and facilitate spectral comparisons.

Table 1. Spectrum table. Peak-level descriptors (mean m/z, standard deviation, AUC) and data quality metrics (r^2, number of signals, confidence intervals) assess Gaussian fit and signal reliability. Confidence intervals support statistical comparison. The table is transposed for clarity, displaying descriptors as rows for easier visualization and interpretation.

	Fragment_1	Fragment_2	...
mz(Da)			
mz_std(Da)			
Int			
Gauss_r2			
N_signals			
Confidence_interval(Da)			
Confidence_interval(ppm)			

mz_min(Da)			
mz_max(Da)			
RelativeIntensity(%)			

This section focuses on establishing a standardized structure for MS2 spectra. The ms2Gauss workflow prioritizes MS2 data analysis to simplify complex sample studies while ensuring consistent data processing, improving comparability and reproducibility across experiments. Prioritizing MS2 over MS1 reduces computational demand while emphasizing structurally informative data (Nothias et al., 2020). Additionally, because adducts are less likely to undergo fragmentation in data-dependent acquisition (DDA) mode, focusing on MS2 helps mitigate their interference in spectral interpretation (Guo & Huan, 2020). To systematically organize MS2-derived information, the All_MS2Samples function iterates through datasets, generating structured tables and metadata to focus on compounds with elucidatable structures. For each sample, All_ms2_spectra processes all MS2 data, outputting a folder of Spectrum tables, while AllMS2Data compiles structured MS2 information into SummMS2 (Table 2.), a summary table integrating metadata such as retention time and retrieval identifiers for streamlined access.

Table 2. SummMS2 table. It stores metadata for each MS2 spectrum, including a unique identifier, precursor ion m/z, and RT.

mz(Da)	RT(s)	id	maxInt
Precursor_1			
Precursor_2			
...			

Resolving MS2 Redundancy and Integrating MS1 Data for an MS2-Based Feature Table

The ms2Gauss workflow ensures high-quality data propagation from raw signals to final conclusions by structuring spectral features as well-defined entities. A feature represents a cluster of signals originating from the same chemical entity, providing a standardized framework for aligning chemical data across samples and conditions (Renner & Reuschenbach, 2022). In ms2Gauss, defining a feature requires precise m/z estimation and chromatographic resolution to ensure distinct clustering of raw signals. The feat_ms2_Gauss function constructs features by eliminating redundancy in SummMS2, refining MS1 descriptors, and structuring processed features into the MS2_Features table (Table 3.), where mass spectrometry and chromatography descriptors are stored in columns. Like SummMS2, MS2_Features includes only features associated with MS2 spectra, ensuring the dataset remains focused on elucidatable compounds. Unlike conventional HRMS workflows that collapse features into a single m/z centroid and retention time (RT) values (Renner & Reuschenbach, 2022), ms2Gauss preserves their signal distribution, enhancing reliability and interpretability.

Table 3. MS2_Features table. Transposed for clarity, this table lists chromatographic (min/max/peak RT) and MS descriptors (m/z, SD, intensity, R², confidence interval), along with internal MS1/MS2 identifiers. Intensity reflects the integrated signal from a single MS1 scan.

	Feature_1	Feature_2	...
ms2_id			
ms1_id			
RT_(s)			
mz_(Da)			
mz_std_(Da)			
I_tol_1spec			
Gauss_r2			
N_points_1spec			
ConfidenceInterval_(Da)			
ConfidenceInterval_(ppm)			
min_mz_(Da)			
max_mz_(Da)			
min_RT_(s)			
max_RT_(s)			
N_ms2_spec			
spectra_id			

Each MS2-producing chemical entity should correspond to a single feature in the feature table. However, in Data-Dependent Acquisition (DDA) mode, high-abundance ions are frequently selected for fragmentation across multiple MS1 scans, generating redundant MS2 spectra with slight RT variations (Guo & Huan, 2020; Pakkir-Shaa et al., 2025). Consequently, SummMS2 includes redundant precursor entries, decreasing computational efficiency and compromising data quality. To address this, ms2Gauss initiates feature construction by processing SummMS2 metadata and removing redundancies using the ms2_SpectralRedundancy function. ms2_SpectralRedundancy groups similar SummMS2 entries into distinct clusters (Figure 2.), ensuring each precursor is uniquely represented. Cluster summary descriptors then serve as a foundation for downstream feature definition.

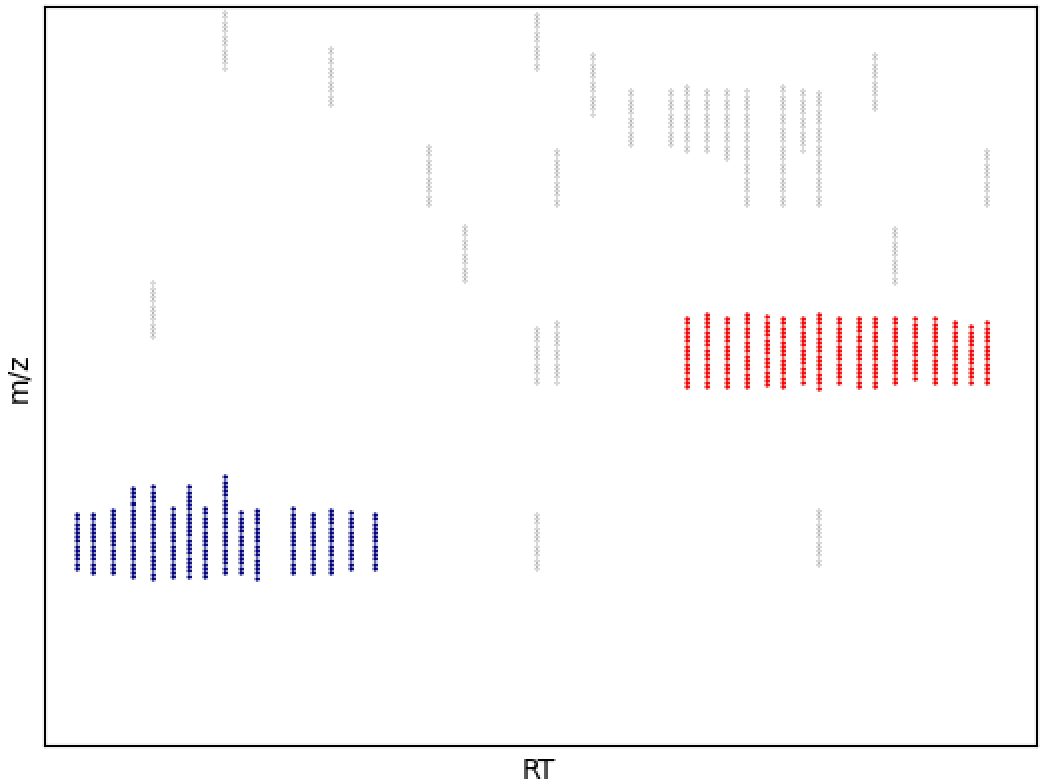

Figure 2. Clustering MS2 Precursors Based on m/z and Retention Time Proximity. This scatter plot visualizes MS2 precursor clustering, with m/z plotted against retention time. Colored clusters highlight grouped precursors exhibiting high intra-cluster similarity, reducing redundancy in feature extraction. Gray points represent unclustered data.

The ms2_SpectralRedundancy function clusters MS2 precursors at two levels, based on proximity and spectral similarity. It operates through ms2_SpectralSimilarityClustering, which constructs an adjacency network linking MS2 precursors based on their m/z and RT proximity, producing preliminary clusters. To further refine clustering, ms2_FeaturesDifferences builds a secondary adjacency network within each cluster derived from ms2_SpectralSimilarityClustering, this time linking nodes based on a cosine similarity threshold between MS2 spectra. This ensures that all resulting features correspond to a unique and representative spectrum, improving reliability in MS2-based analysis.

Feature tables are crucial for sample alignment and other data-driven analyses requiring high-quality data (Renner & Reuschenbach, 2022). Resolving redundancy in SummMS2 enhances one-to-one feature matching, minimizing alignment errors and improving comparability. However, alignment accuracy still depends on the integrity of individual features (Nothias et al., 2020). By integrating MS1 data, ms2Gauss refines precursor m/z reliability, improves RT consistency, and quantifies uncertainty, strengthening feature matching within the limits of data variability.

In ms2Gauss, feat_ms2_Gauss calls ms2_features_stats, which then triggers closest_ms1_spec to refine MS1 data by selecting MS1 peaks with the closest m/z and RT to the SummMS2 precursor metadata. To ensure statistical rigor, ms2_features_stats retains Gaussian descriptors from the closest precursor peak across at least three MS1 scans. Also inside feat_ms2_Gauss, the mzPeak function processes MS1 peaks similarly to how ms2Peak handles MS2 peaks. At this stage, the feature table indicates only feature presence or absence. For a higher computational cost, ms2Gauss enables quantitative refinement by extracting and analyzing full chromatograms for each feature.

The ms2Gauss workflow refines SummMS2 metadata by clustering redundant MS2 spectra, integrating MS1 data, and constructing a structured MS2_Features table. This feature table enhances HRMS analysis by integrating confidence intervals and other quality metrics, which characterize signal distribution and feature reliability. A structured data framework is essential, as raw spectral data is inherently noisy and sparse (Renner & Reuschenbach, 2022). A well-structured MS2_Features table consolidates essential information, enabling robust cross-sample comparisons, precise statistical analyses, and reproducible data-driven modeling. The structured design of MS2_Features ensures seamless data integration and compatibility with other HRMS workflows and analytical pipelines.

Align, Filter, and Refine: Identifying Features in Experimental Data

The AlignedSamplesDF integrates multiple MS2-based feature tables from different samples, increasing confidence in each feature, which, once in AlignedSamplesDF (Table 4.), becomes a consensus across all aligned samples. A feature that can be explained is more likely to be valid, reproducible, and meaningful. A feature that appears consistently across samples but lacks interpretability may be statistically reliable, yet without chemical or biological relevance, underscoring the necessity of integrating statistical and experimental validation for meaningful feature assessment (Renner & Reuschenbach, 2022). While aligning samples reveals inherent patterns, filtering these patterns based on experimental design sharpens the dataset, yielding CarbonSourceFeatures, which retain experimental relevance. Additionally, analyzing the entire chromatogram for the remaining features after filtering ensures that further analysis with the CarbonSourceFeatures table (Table 5.) remains quantitative while minimizing computational demand.

Table 4. AlignedSamplesDF: A Qualitative Framework for Tracking Feature Presence and Alignment in HRMS

mz_(Da)	mz_std_(Da)	RT_(s)	Sample_1	Sample_2	Sample_3	...
Feature_1			1	1	0	...
Feature_2			0	1	1	...
Feature_3			0	1	1	...
...			...	...	...	...

Table 5. CarbonSourceFeatures: A Qualitative Table for Filtering Biologically Relevant Features Based on Carbon Source and Experimental Design in HRMS Analysis.

mz_(Da)	mz_std_(Da)	RT_(s)	Sample_1	Sample_3	...
Feature_1			1714.3	0	...
Feature_2			0	2599.15	...

Feature_3			0	603.843	...
...			...	...	...

The Feature_ms2_SamplesAlignment function enhances the consistency, comparability, and reliability of MS2-derived features across samples while improving feature identification through contextual alignment in the resulting AlignedSamplesDF. By supporting iterative refinements in experimental workflows, the structured data management in ms2Gauss also enhances hypothesis validation and data interpretation. Consensus features emerge from distinctive clusters, where clustering validates feature identity by detecting consistent patterns across all datasets (Pakkir-Shaa et al., 2025). This structured approach improves data organization, enabling efficient querying, feedback, and automation.

The Feature_ms2_SamplesAlignment function clusters related features based on _m/z_, retention time (RT) proximity, and MS2 spectral similarity. Unlike ms2_SpectralRedundancy, which processes a single sample, Feature_ms2_SamplesAlignment extends clustering across multiple datasets. By directly comparing _m/z_ and MS2 spectra, this approach ensures reliable feature alignment regardless of chromatographic method differences. MS2 spectral data play a key role in structural matching, compensating for retention time variations caused by chromatographic conditions, matrix effects, and sample processing differences. While _m/z_ remains stable for cross-sample comparisons, MS2 spectra provide additional confidence in feature identity despite sample-dependent variability.

Cluster analysis organizes similar features, transforming each sample from an imprecise vector in feature space into structured columns in a consensus features table. This transition preserves the advantages of mathematical operations, enhancing accuracy on downstream analyses. Statistical data processing further refines this alignment by standardizing descriptors, mitigating sample variability, and improving data reliability (Heuckeroth et al., 2024). These refinements ensure robust data interpretation, facilitating cross-sample consistency and enabling more precise downstream analysis. However, a features table alone lacks meaning—proper experimental design and context are essential for reliable interpretation, especially in non-target analyses, where most detected substances remain poorly studied (Renner & Reuschenbach, 2022).

Accurate data analysis depends on a well-structured experimental framework, and ms2Gauss provides guidance for integrating relevant experimental data to ensure reliable interpretation. The experiment-specific functions RemoveBlankFeatures and RefineFeatureTable_withChromatogram leverage experimental data to contrast samples and refine features, respectively. These functions incorporate biological context by analyzing diverse sample types, refining feature identification, and improving metabolomics analysis through comparative evaluation.

Our metabolomics experiment examined bacterial metabolism of selected pharmaceutical drugs and its variability under different primary carbon sources. A microbial community from activated sludge was cultivated in a retentostat system under distinct carbon-source conditions to ensure metabolic stability. The reactor feed was then spiked with a pharmaceutical drug mixture, and samples were collected for chemical analysis. A total of 54 samples, including biological and technical replicates, blanks, and controls, were analyzed using an Orbitrap™ high-resolution mass spectrometer. The resulting MS2-based feature table, generated using Feature_ms2_SamplesAlignment, aligned these samples and identified 1571 distinct features for downstream filtering and refinement.

Only a small subset of the 1571 detected features corresponded to bacterial transformation products of pharmaceutical drugs, requiring additional filtering to assess the influence of the carbon source. Within RemoveBlankFeatures, Samples_NFeatures_Filter refines AlignedSamplesDF by selecting features based on sample appearance, leveraging attributes from the experimental attributes table (Table 6.). This function ensures experimental reproducibility by contrasting samples expected to differ while grouping similar ones. To isolate HRMS features associated with pharmaceutical drug biotransformation, we analyzed various sample types, including effluents from bioreactors treating pharmaceutical drugs, effluents from bioreactors without pharmaceutical drugs (negative controls), and influent samples with and without pharmaceutical drugs to establish baseline chemical compositions. The majority of detected features (52.9%) were present in effluent samples treating pharmaceutical drugs, whereas possible transformation products unique to these sample types comprised only 19.7% (Figure 3.). By removing features found in blanks and controls, we retained only those specific to pharmaceutical drug metabolism, ensuring a focused and biologically relevant dataset.

Table 6. Experimental attributes table: Sample type and carbon source as key experimental variables.

id	Origin	Experiment	SampleName
Sample_1			Sample_1_id+.mzML
Sample_2			
...			

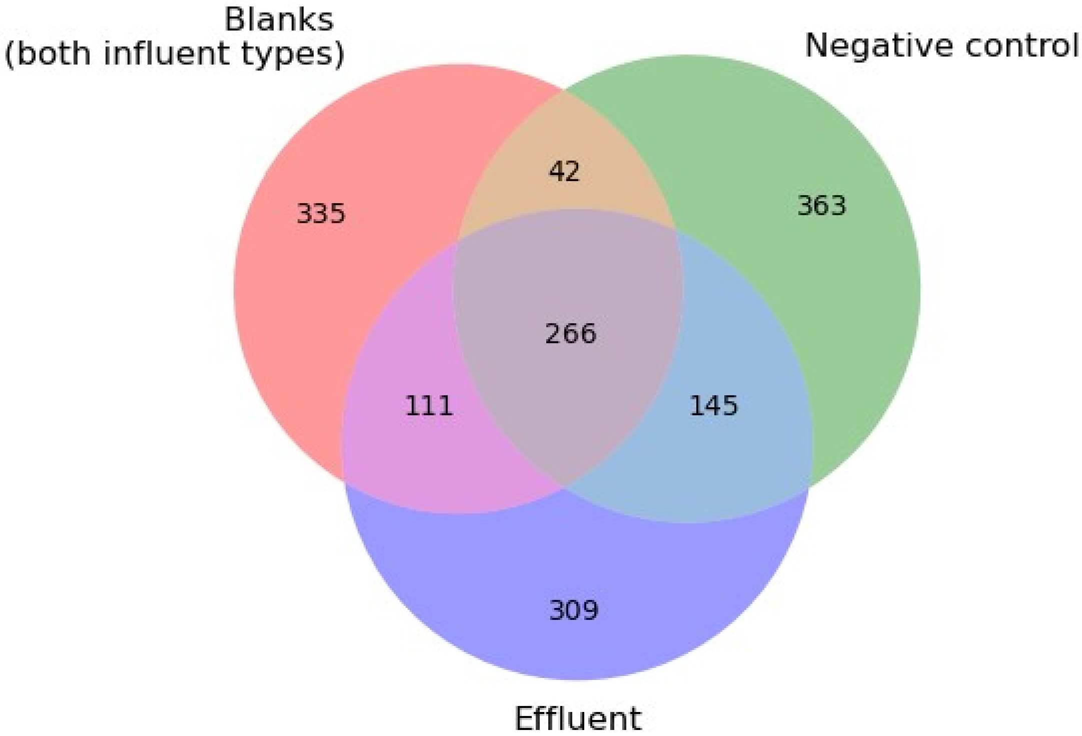

Figure 3. Filtering and contrasting sample types. Distribution of detected features across Blanks, Negative Control, and Effluent samples. 266 features are shared across all three conditions, while 335, 363, and 309 are unique to Blanks, Negative Control, and Effluent, respectively.

While chromatographic analysis is computationally demanding, its cost remains minimal in this case, as the number of features undergoing further refinement is limited. ms2Gauss enhances the reliability of the CarbonSourceFeatures table by refining the corresponding high-resolution extracted ion chromatogram (HR-EIC) through deconvolution, peak identification, and quantification using the area under the curve (AUC). The RefineFeatureTable_withChromatogram function systematically propagates MS2 features across samples by leveraging MS1 data, consolidating fragmentation data from replicates and independent samples to create a comprehensive feature representation.

Within the RefineFeatureTable_withChromatogram function, ExtractAllRawPeaks searches for raw peaks across all spectra within predefined _m/z_ and RT thresholds for each feature in each sample inside CarbonSourceFeature, and generates the corresponding chromatograms. These chromatograms are processed by ResolveFullChromatogram, which applies Gaussian deconvolution via SciPy's `curve_fit` (Virtanen P. et al., 2020) to decompose chromatograms into individual peaks, expressing the chromatogram as a linear combination of Gaussians (Figure 4.) (Di Marco V. & Bombi G., 2001). The function Match_ms2Feature_Chrom, which is called by RefineFeatureTable_withChromatogram, preserves the resolved peak with the closest RT to the MS2-based feature RT, ensuring precise RT consistency across samples and improving feature detection accuracy and reliability.

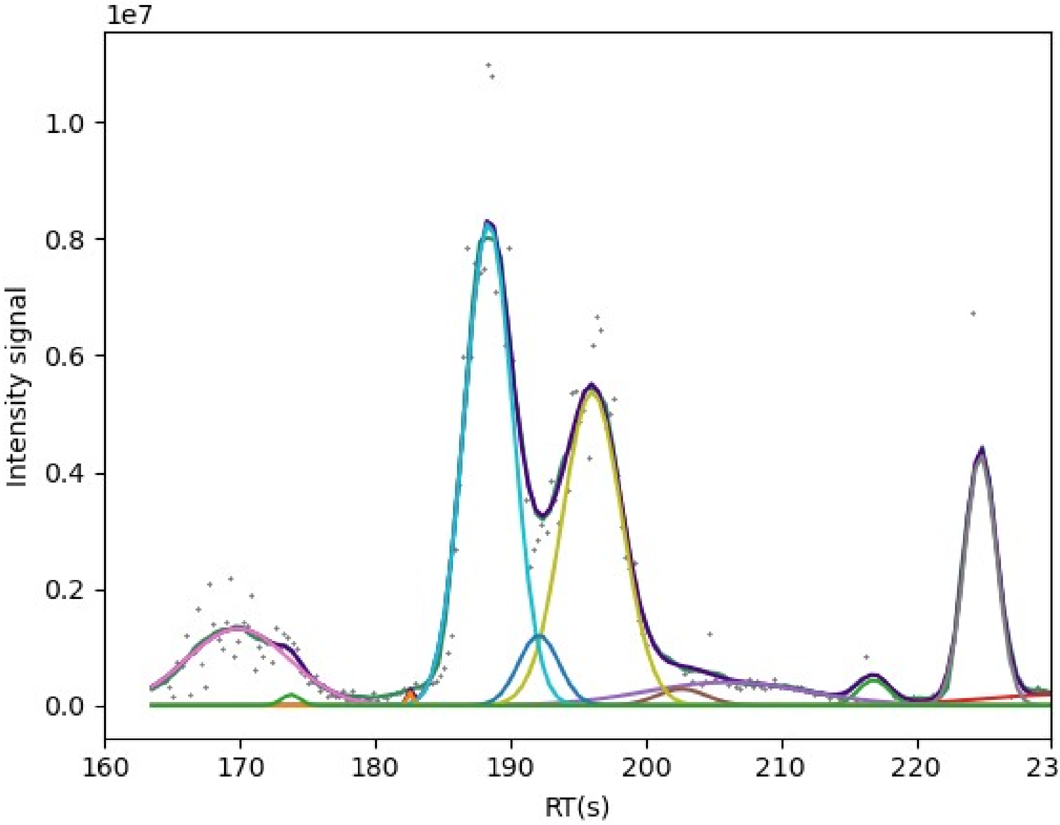

Figure 4. Gaussian mixture model representation of a chromatogram for peak Identification and Characterization.

By systematically integrating statistical analysis, spectral alignment, and chromatographic refinement, ms2Gauss provides a robust and scalable framework for metabolomics research. The resulting data enable reliable feature identification and quantitative comparisons across complex biological systems, supporting hypothesis-driven investigations and enhancing the reproducibility of experimental findings.

References

- Adusumilli, R., & Mallick, P. (2017). Data conversion with proteoWizard msConvert. Methods in Molecular Biology, 1550, 339–368. https://doi.org/10.1007/978-1-4939-6747-6_23/FIGURES/10
- Deutsch, E. W. (2010). Mass Spectrometer Output File Format mzML. Methods in Molecular Biology (Clifton, N.J.), 604, 319. https://doi.org/10.1007/978-1-60761-444-9_22
- di Marco, V. B., & Bombi, G. G. (2001). Mathematical functions for the representation of chromatographic peaks. Journal of Chromatography A, 931(1–2), 1–30. https://doi.org/10.1016/S0021-9673(01)01136-0
- Eliuk, S., & Makarov, A. (2015). Evolution of Orbitrap Mass Spectrometry Instrumentation. Annual Review of Analytical Chemistry (Palo Alto, Calif.), 8, 61–80. https://doi.org/10.1146/ANNUREV-ANCHEM-071114-040325
- Guo, J., & Huan, T. (2020). Comparison of Full-Scan, Data-Dependent, and Data-Independent Acquisition Modes in Liquid Chromatography-Mass Spectrometry Based Untargeted Metabolomics. Analytical Chemistry, 92(12), 8072–8080.

https://doi.org/10.1021/ACS.ANALCHEM.9B05135/SUPPL_FILE/AC9B05135_SI_001.PDF

- Heuckeroth, S., Damiani, T., Smirnov, A., Mokshyna, O., Brungs, C., Korf, A., Smith, J. D., Stincone, P., Dreolin, N., Nothias, L. F., Hyötyläinen, T., Orešič, M., Karst, U., Dorrestein, P. C., Petras, D., Du, X., van der Hooft, J. J. J., Schmid, R., & Pluskal, T. (2024). Reproducible mass spectrometry data processing and compound annotation in MZmine 3. Nature Protocols 2024 19:9, 19(9), 2597–2641. https://doi.org/10.1038/s41596-024-00996-y
- Nothias, L. F., Petras, D., Schmid, R., Dührkop, K., Rainer, J., Sarvepalli, A., Protsyuk, I., Ernst, M., Tsugawa, H., Fleischauer, M., Aicheler, F., Aksenov, A. A., Alka, O., Allard, P. M., Barsch, A., Cachet, X., Caraballo-Rodriguez, A. M., da Silva, R. R., Dang, T., … Dorrestein, P. C. (2020). Feature-based molecular networking in the GNPS analysis environment. Nature Methods 2020 17:9, 17(9), 905–908. https://doi.org/10.1038/s41592-020-0933-6
- Pakkir Shah, A. K., Walter, A., Ottosson, F., Russo, F., Navarro-Diaz, M., Boldt, J., Kalinski, J. C. J., Kontou, E. E., Elofson, J., Polyzois, A., González-Marín, C., Farrell, S., Aggerbeck, M. R., Pruksatrakul, T., Chan, N., Wang, Y., Pöchhacker, M., Brungs, C., Cámara, B., … Petras, D. (2024). Statistical analysis of feature-based molecular networking results from non-targeted metabolomics data. Nature Protocols 2024 20:1, 20(1), 92–162. https://doi.org/10.1038/s41596-024-01046-3
- Perry, R. H., Cooks, R. G., & Noll, R. J. (2008). Orbitrap mass spectrometry: Instrumentation, ion motion and applications. Mass Spectrometry Reviews, 27(6), 661–699. https://doi.org/10.1002/MAS.20186
- Renner, G., & Reuschenbach, M. (2023). Critical review on data processing algorithms in non-target screening: challenges and opportunities to improve result comparability. Analytical and Bioanalytical Chemistry, 415(18), 4111–4123. https://doi.org/10.1007/S00216-023-04776-7/TABLES/1
- Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., … Vázquez-Baeza, Y. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. Nature Methods 2020 17:3, 17(3), 261–272. https://doi.org/10.1038/s41592-019-0686-2