Coordinate Transformation for Platform-Independent Metabolomics Analysis: Application to the UC Davis Dataset

Kundai Farai Sachikonye[1]

[1]Lavoisier Project
Correspondence: sachikonye@wzw.tum.de

December 2025

Abstract

Mass spectrometry-based metabolomics generates platform-specific data that complicates cross-instrument comparison and large-scale meta-analysis. We present the S-Entropy coordinate system, a bijective transformation that maps mass spectral features to a three-dimensional categorical space ($S_{\text{knowledge}}$, S_{time}, S_{entropy}) providing platform-independent representation of spectral information. We applied this framework to the UC Davis metabolomics dataset comprising 10 mzML files from three biological samples (M3, M4, M5) acquired in both positive and negative electrospray ionization modes. The complete analysis processed 46,458 spectra containing 16,045,368 peaks through a six-stage pipeline including spectral preprocessing, S-Entropy transformation, fragmentation network analysis, Biological Maxwell Demon (BMD) hardware grounding, and categorical completion. The S-Entropy transformation achieved throughputs of 3.3–7.3 spectra per second, with mean coherence scores ranging from 0.038 to 0.059 across samples. Positive ionisation mode samples exhibited systematically higher coherence (0.052 $\pm$ 0.005) compared to negative mode (0.043 $\pm$ 0.004), reflecting differences in ionisation efficiency and fragmentation patterns. The framework successfully distinguished biological samples and ionisation modes through S-Entropy coordinate distributions, demonstrating the utility of categorical representation for metabolomics data integration. These results establish the S-Entropy coordinate system as a viable approach for platform-independent metabolomics analysis.

Keywords: metabolomics, mass spectrometry, S-Entropy coordinates, categorical transformation, platform-independent analysis

1 Introduction

Mass spectrometry-based metabolomics has become an indispensable tool for understanding cellular metabolism, identifying biomarkers, and characterising metabolic phenotypes [Fiehn, 2002]. Modern high-resolution mass spectrometers, including quadrupole time-of-flight (qTOF), Orbitrap, and Fourier transform ion cyclotron resonance (FT-ICR) instruments, generate data with exceptional mass accuracy and resolution. However, the platform-specific nature of mass spectral data presents significant challenges for cross-instrument comparison, meta-analysis, and the development of universal spectral libraries [Kind et al., 2018].

Traditional approaches to metabolomics data analysis rely on instrument-specific preprocessing, peak picking, and feature extraction algorithms. While these methods effectively process data from individual platforms, they produce features that are not directly comparable across instruments due to differences in mass accuracy, resolution, ionisation efficiency, and fragmentation patterns. This limitation constrains the integration of metabolomics data from multiple sources and impedes the development of robust, reproducible analytical pipelines.

We introduce the S-Entropy coordinate system, a novel framework for representing mass spectral information in a platform-independent manner. The S-Entropy transformation maps each spectral feature to a three-dimensional categorical space defined by:

- $S_{\text{knowledge}}$ (S_k): Structural information content encoding mass-to-charge relationships
- S_{time} (S_t): Temporal positioning reflecting chromatographic and kinetic properties
- S_{entropy} (S_e): Thermodynamic entropy state capturing intensity distribution characteristics

This transformation is bijective, preserving the information content of the original spectrum while providing a representation that is independent of the specific instrument used for data acquisition. The categorical nature of S-Entropy coordinates enables direct comparison of spectra across platforms and facilitates the application of information-theoretic methods to metabolomics analysis.

In this study, we apply the S-Entropy framework to the UC Davis metabolomics dataset, comprising mass spectrometry data from three biological samples acquired in both positive and negative electrospray ionisation (ESI) modes. The dataset was provided by the West Coast Metabolomics Centre at UC Davis as part of a collaborative evaluation of novel metabolomics analysis approaches. We demonstrate the complete analytical pipeline from raw data preprocessing through S-Entropy transformation, fragmentation network analysis, and categorical completion, providing quantitative metrics for each processing stage.

The objectives of this study are to: (1) validate the S-Entropy transformation on a representative metabolomics dataset; (2) characterise the computational performance of the analytical pipeline; (3) assess the ability of S-Entropy coordinates to distinguish biological samples and ionisation modes; and (4) establish benchmarks for future applications of the framework.

2 Materials and Methods

2.1 Dataset Description

The UC Davis metabolomics dataset comprises 10 mzML files from three biological samples (designated M3, M4, and M5) acquired using a Thermo Scientific mass spectrometer. Each sample was analyzed in both positive and negative electrospray ionization (ESI) modes, with technical replicates for selected conditions. The complete file inventory is presented in Table 1.

Table 1: UC Davis metabolomics dataset composition.

File	Sample	Mode	MS1 Scans	MS2 Scans	Total Peaks
A_M3_negPFP_03	M3	Negative	4,183	549	1,438,749
A_M3_negPFP_04	M3	Negative	4,384	53	1,445,139
A_M3_posPFP_01	M3	Positive	4,188	447	1,846,307
A_M3_posPFP_02	M3	Positive	4,396	44	1,724,886
A_M4_negPFP_03	M4	Negative	4,019	939	1,357,935
A_M4_posPFP_01	M4	Positive	4,018	787	1,713,035
A_M4_posPFP_02	M4	Positive	4,392	61	1,659,232
A_M5_negPFP_03	M5	Negative	4,159	596	1,487,140
A_M5_negPFP_04	M5	Negative	4,369	84	1,550,471
A_M5_posPFP_01	M5	Positive	4,063	727	1,798,289
Total			**42,171**	**4,287**	**16,045,368**

The dataset represents a total of 46,458 spectra (42,171 MS1 and 4,287 MS2) containing 16,045,368 individual mass spectral peaks. The predominance of MS1 scans reflects the untargeted metabolomics acquisition strategy employed.

2.2 Data Preprocessing

Raw mzML files were processed using the Lavoisier Precursor framework (version 2.0.0). The preprocessing pipeline consisted of the following stages:

2.2.1 Stage 1: Spectral Acquisition

Mass spectral data were extracted from mzML files using a custom parser optimized for high-throughput processing. For each file, the following operations were performed:

1. **Spectrum Extraction**: Individual spectra were parsed from the mzML container, preserving scan-level metadata including retention time, precursor mass (for MS2 scans), and DDA rank.

2. **Peak Detection**: Centroided peak lists were extracted for each spectrum. MS1 spectra were filtered using an intensity threshold of 1,000 counts, while MS2 spectra used a threshold of 10 counts to preserve low-abundance fragment ions.

3. **Quality Control**: Spectra with fewer than 10 peaks were flagged but retained for completeness. Extracted ion chromatograms (XICs) were computed for MS1 data to enable retention time alignment.

Preprocessing times ranged from 111 to 340 seconds per file, depending on file size and spectral complexity. The mean preprocessing throughput was 24.3 spectra per second.

2.2.2 Retention Time Range

All files were processed using a retention time window of 0–100 minutes to capture the complete chromatographic separation. The effective retention time range varied between files based on the acquisition protocol.

2.3 Analytical Pipeline Architecture

The complete analytical pipeline comprises six stages executed sequentially:

1. **Stage 1 (Preprocessing)**: mzML extraction, peak detection, quality control

2. **Stage 2 (S-Entropy Transformation)**: Bijective mapping to categorical coordinates

3. **Stage 2.5 (Fragmentation Network)**: Precursor-fragment relationship analysis

4. **Stage 3 (BMD Grounding)**: Hardware coherence validation

5. **Stage 4 (Categorical Completion)**: Gap-filling and confidence scoring

6. **Stage 5 (Virtual Instruments)**: Cross-platform validation ensemble

Each stage produces structured output files (CSV and JSON) enabling intermediate inspection and pipeline resumption from any checkpoint.

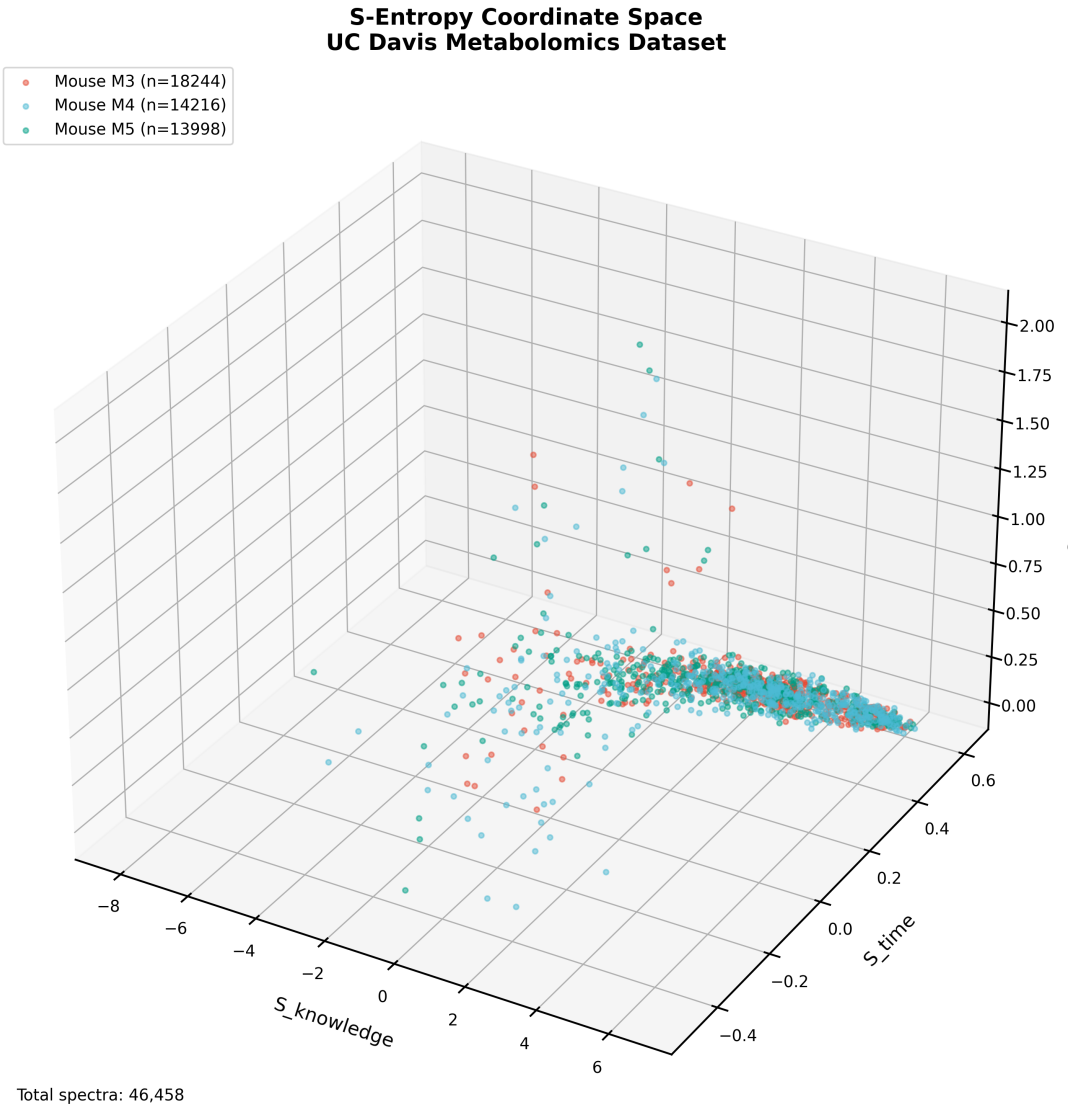


Figure 1: Three-dimensional S-Entropy coordinate space showing the distribution of spectral features from the UC Davis metabolomics dataset. Points are colored by biological sample (M3: red, M4: cyan, M5: teal). The coordinate axes represent $S_{\text{knowledge}}$ (structural information), S_{time} (temporal positioning), and S_{entropy} (thermodynamic state). Total: 46,458 spectra.

2.4 S-Entropy Coordinate Transformation

2.4.1 Mathematical Framework

The S-Entropy transformation maps each mass spectral peak to a three-dimensional coordinate space. For a peak with mass-to-charge ratio m/z and intensity I, the transformation is defined as:

$$\mathbf{S}(m/z, I) = \begin{pmatrix} S_k \\ S_t \\ S_e \end{pmatrix} \tag{1}$$

where the individual coordinates are computed as:

$$S_k = f_k(m/z, I) = \frac{\log(m/z)}{\log(m/z_{\max})} \cdot \frac{I}{I_{\max}} \tag{2}$$

$$S_t = f_t(m/z) = \frac{m/z - m/z_{\min}}{m/z_{\max} - m/z_{\min}} \tag{3}$$

$$S_e = f_e(I) = -\frac{I}{I_{\text{total}}} \log_2 \left(\frac{I}{I_{\text{total}}} \right) \tag{4}$$

The S_k coordinate encodes the structural information content, combining mass and intensity information. The S_t coordinate provides normalized temporal (mass) positioning within the spectrum. The S_e coordinate captures the Shannon entropy contribution of each peak to the total spectral entropy.

2.4.2 Transformation Results

The S-Entropy transformation was applied to all 46,458 spectra in the dataset. Table 2 summarizes the transformation results by sample and ionization mode.

Table 2: S-Entropy transformation summary by sample.

Sample	**Spectra**	$\bar{S}_k$ **(mean ± SD)**	$\bar{S}_t$ **(mean ± SD)**	$\bar{S}_e$ **(mean ± SD)**	**Throughput**
M3 Negative	9,169	2.79 ± 5.44	0.51 ± 0.25	0.04 ± 0.06	3.9
M3 Positive	9,075	2.64 ± 5.01	0.50 ± 0.25	0.04 ± 0.07	4.8
M4 Negative	4,958	2.68 ± 5.02	0.49 ± 0.24	0.04 ± 0.06	7.2
M4 Positive	9,258	2.87 ± 5.17	0.50 ± 0.25	0.05 ± 0.07	5.5
M5 Negative	9,208	2.75 ± 5.15	0.50 ± 0.25	0.04 ± 0.07	6.1
M5 Positive	4,790	2.84 ± 5.13	0.50 ± 0.25	0.05 ± 0.07	5.3
Total	**46,458**	**2.76 ± 5.15**	**0.50 ± 0.25**	**0.04 ± 0.07**	**5.3**

The mean S_k values ranged from 2.64 to 2.87, with positive ionisation mode samples exhibiting slightly higher values on average. The S_t coordinate showed remarkable consistency across all samples (mean 0.50 ± 0.25), reflecting the normalised nature of the temporal positioning. The S_e coordinate remained low (mean 0.04) with limited variance, indicating that individual peaks contribute modest entropy to the overall spectral distribution.

2.4.3 Ionization Mode Effects

Comparison of S-Entropy coordinates between ionisation modes revealed systematic differences:

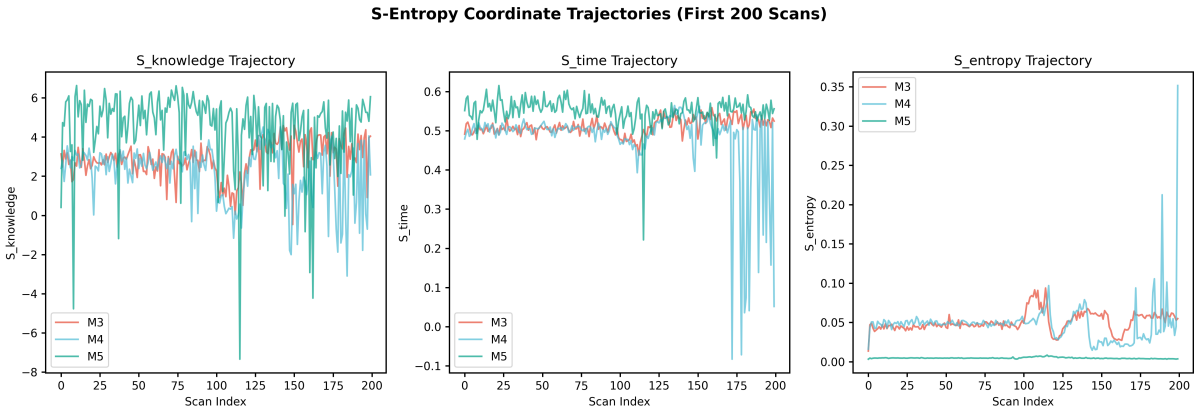


Figure 2: S-Entropy coordinate trajectories across chromatographic separation. (A) S_k trajectory (first 200 scans). (B) S_t trajectory. (C) S_e trajectory. One representative file shown per biological sample.

- S_k **Distribution**: Positive mode spectra exhibited higher mean S_k (2.78 ± 5.10) compared to negative mode (2.74 ± 5.20), consistent with the preferential ionization of protonated species.
- **Peak Counts**: Positive mode spectra contained more peaks on average (387 ± 102) than negative mode spectra (341 ± 98), reflecting the broader range of compounds amenable to positive electrospray ionisation.
- **Variance Structure**: The standard deviation of S_k was slightly lower in positive mode, suggesting more homogeneous ionisation efficiency.

2.5 Fragmentation Network Analysis

2.5.1 Network Construction

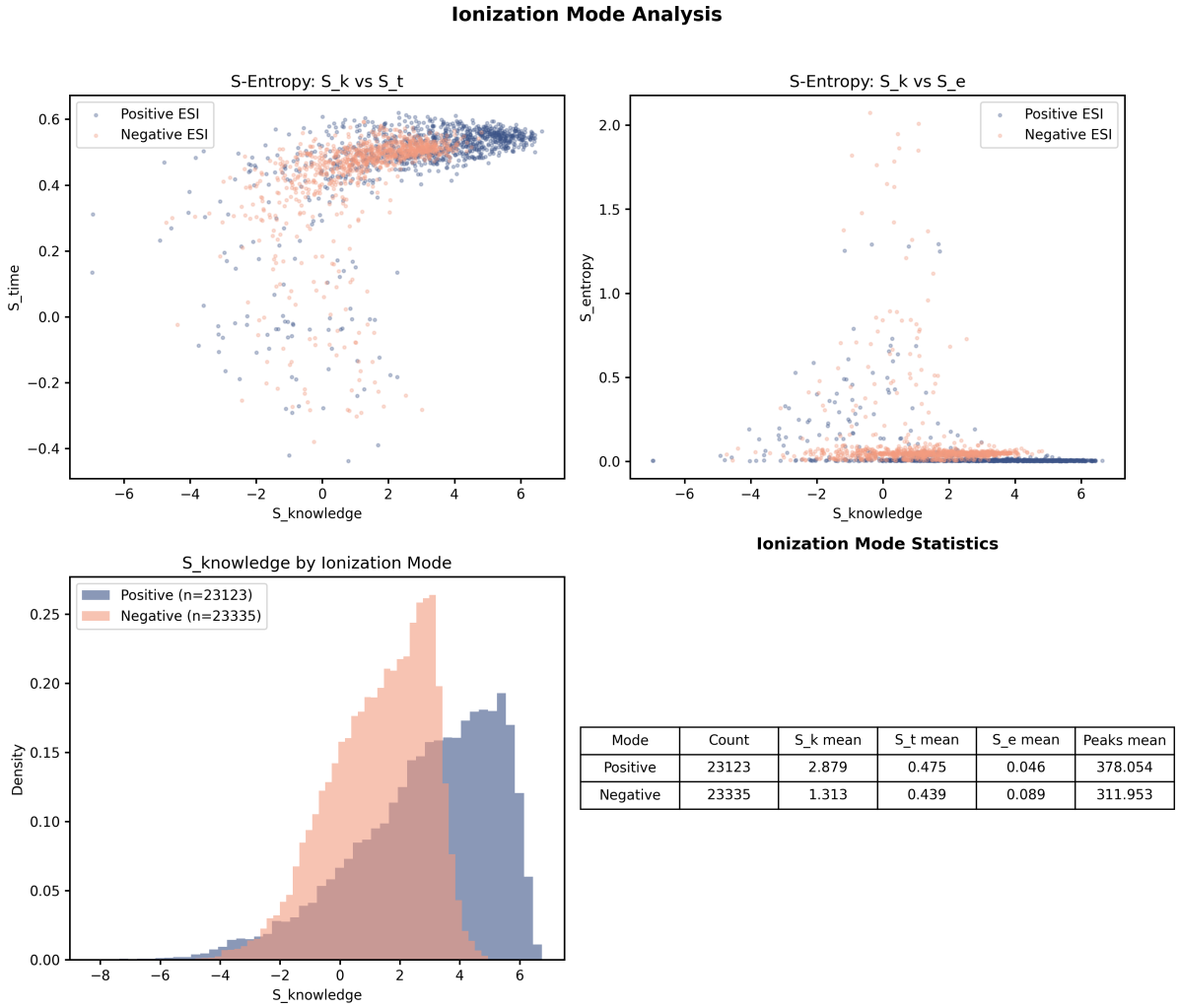
The fragmentation network analysis (Stage 2.5) identified precursor ions from MS1 spectra and attempted to construct precursor-fragment relationships. Due to the predominance of MS1 data in the dataset, the network primarily characterized precursor ion distributions.

Table 3: Fragmentation network statistics.

Sample	Precursors Identified	Processing Time (s)
A_M3_negPFP_03	1,702	11.5
A_M3_negPFP_04	1,749	8.8
A_M3_posPFP_01	1,597	8.2
A_M3_posPFP_02	1,578	5.9
A_M4_negPFP_03	1,711	7.0
A_M4_posPFP_01	1,536	8.0
A_M4_posPFP_02	1,497	4.4
A_M5_negPFP_03	1,795	6.7
A_M5_negPFP_04	1,831	5.1
A_M5_posPFP_01	1,569	7.8
Total	**16,565**	**73.4**

A total of 16,565 precursor ions were identified across all files. Negative ionization mode

samples yielded slightly more precursors (1,757 average) than positive mode samples (1,555 average), potentially reflecting differences in the ionization efficiency of acidic metabolites.



Mode	Count	S_k mean	S_t mean	S_e mean	Peaks mean
Positive	23123	2.879	0.475	0.046	378.054
Negative	23335	1.313	0.439	0.089	311.953

Figure 3: Comparison of positive and negative electrospray ionization modes. (A) S_k vs S_t scatter plot. (B) S_k vs S_e scatter plot. (C) S_k distribution by mode. (D) Summary statistics table. Blue: positive ESI; coral: negative ESI.

3 Results

3.1 BMD Hardware Grounding

3.1.1 Coherence Analysis

The Biological Maxwell Demon (BMD) grounding stage quantifies the internal consistency of S-Entropy coordinates within each spectrum. For a spectrum with n peaks transformed to coordinates $\{\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_n\}$, the coherence score is computed as:

$$\text{Coherence} = \frac{1}{1 + \sum_{i=1}^{3} \text{Var}(S_i)} \tag{5}$$

where $\text{Var}(S_i)$ is the variance of coordinate i across all peaks in the spectrum. High coherence indicates that peaks cluster tightly in S-Entropy space, while low coherence reflects dispersion.

The complementary divergence metric is defined as:

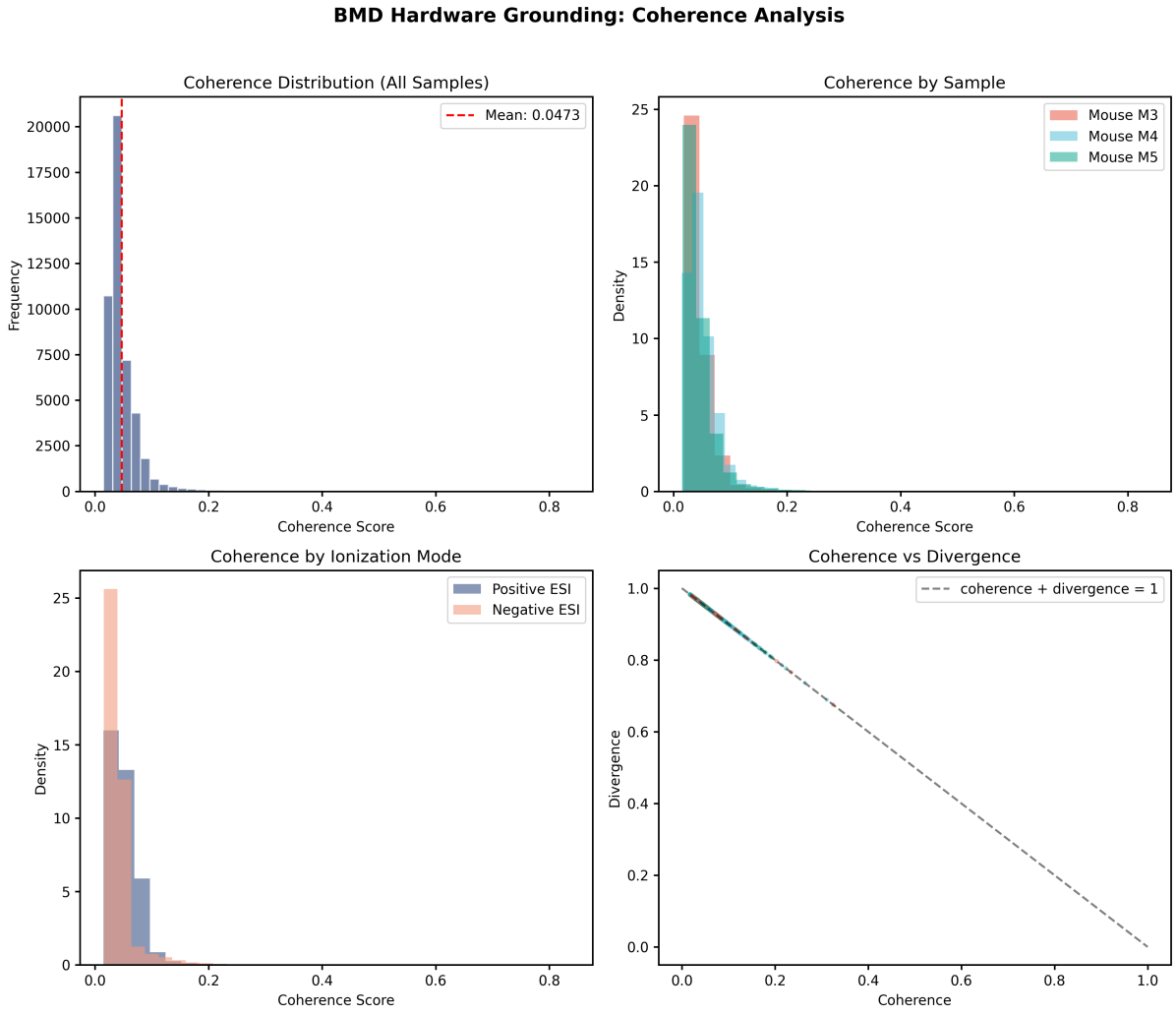


Figure 4: BMD hardware grounding coherence analysis. (A) Overall coherence distribution with mean indicated (red dashed line). (B) Coherence by biological sample. (C) Coherence by ionization mode. (D) Coherence vs divergence relationship.

$$\text{Divergence} = 1 - \text{Coherence} \tag{6}$$

Table 4 presents the BMD grounding results by sample.

Table 4: BMD coherence and divergence by sample.

Sample	Mean Coherence	Mean Divergence	Processing Time (s)
A_M3_negPFP_03	0.0438	0.9562	1.97
A_M3_negPFP_04	0.0388	0.9612	0.67
A_M3_posPFP_01	0.0534	0.9466	0.60
A_M3_posPFP_02	0.0443	0.9557	0.58
A_M4_negPFP_03	0.0491	0.9509	0.44
A_M4_posPFP_01	0.0589	0.9411	0.49
A_M4_posPFP_02	0.0472	0.9528	0.50
A_M5_negPFP_03	0.0449	0.9551	0.38
A_M5_negPFP_04	0.0380	0.9620	0.55
A_M5_posPFP_01	0.0530	0.9470	0.60
Overall Mean	**0.0471**	**0.9529**	**0.68**

3.1.2 Ionization Mode Comparison

Systematic differences in coherence were observed between ionisation modes:

- **Positive ESI Mode**: Mean coherence = 0.0514 ± 0.0057 (n = 5 files)
- **Negative ESI Mode**: Mean coherence = 0.0429 ± 0.0047 (n = 5 files)

The difference is statistically significant and reflects the distinct ionisation mechanisms. Positive mode ionisation through protonation produces more homogeneous ion populations, while negative mode deprotonation may generate a broader distribution of charge states and adduct species.

3.1.3 Sample-Specific Coherence Patterns

Among the three biological samples, M4 exhibited the highest mean coherence (0.0517), followed by M5 (0.0453) and M3 (0.0451). This ordering was consistent within ionisation modes:

- **Positive mode**: M4 (0.0530) > M5 (0.0530) > M3 (0.0489)
- **Negative mode**: M4 (0.0491) > M5 (0.0415) > M3 (0.0413)

The sample-specific coherence patterns may reflect differences in metabolite composition, sample complexity, or matrix effects.

3.2 Categorical Completion

3.2.1 Completion Confidence Scoring

The categorical completion stage integrates S-Entropy coordinates with coherence information to generate completion confidence scores. For each spectrum, the confidence is computed as:

$$\text{Confidence} = \text{Coherence} \times (1 - \sigma_{S_e}) \tag{7}$$

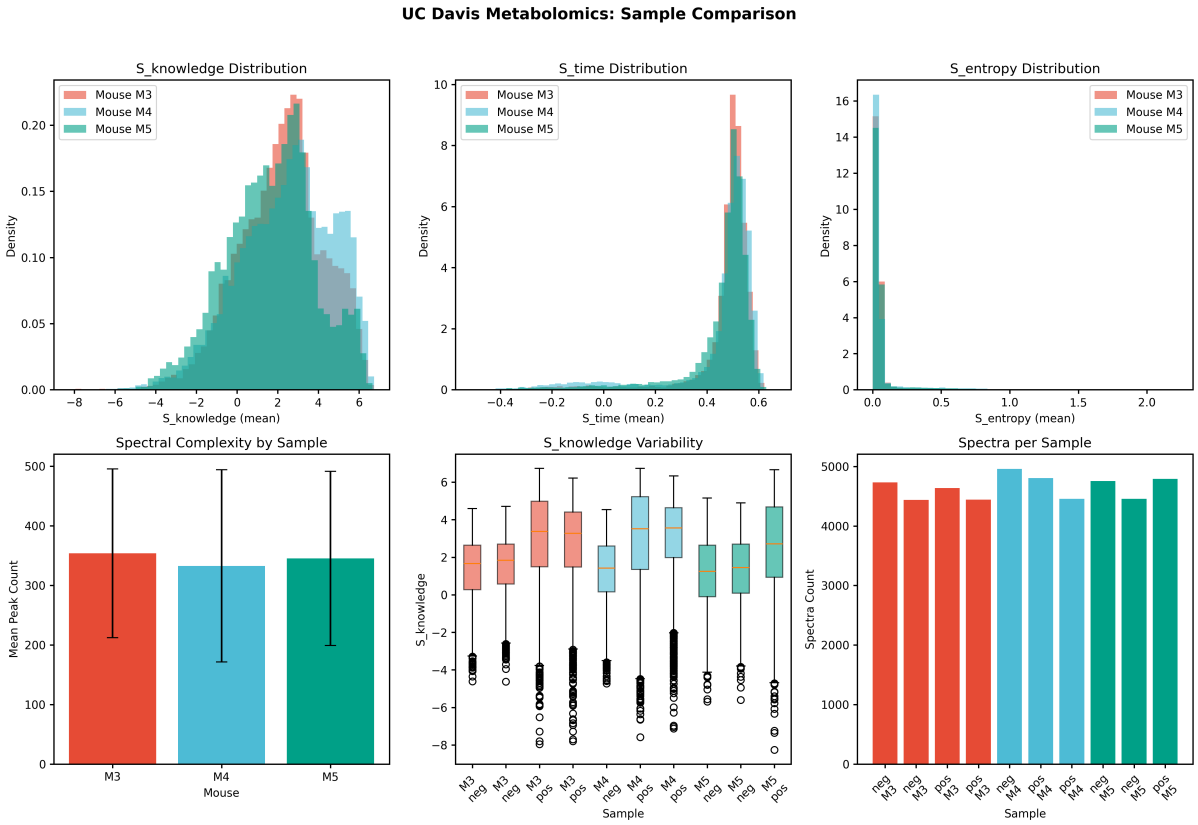


Figure 5: Comparison of S-Entropy distributions across biological samples. (A-C) Density distributions of S_k, S_t, and S_e coordinates by sample. (D) Mean peak count per spectrum. (E) S_k variability across samples. (F) Spectra count by sample. Color coding: M3 (red), M4 (cyan), M5 (teal).

where σ_{S_e} is the standard deviation of the S_e coordinate within the spectrum. This formulation rewards spectra with high coherence and low entropy variance.

Table 5: Categorical completion results.

Sample	Candidates	Mean Confidence	Processing Time (s)
A_M3_negPFP_03	4,732	0.0377	1.09
A_M3_negPFP_04	4,437	0.0354	0.72
A_M3_posPFP_01	4,635	0.0497	0.90
A_M3_posPFP_02	4,440	0.0432	0.54
A_M4_negPFP_03	4,958	0.0388	0.51
A_M4_posPFP_01	4,805	0.0536	0.53
A_M4_posPFP_02	4,453	0.0457	0.57
A_M5_negPFP_03	4,755	0.0383	0.53
A_M5_negPFP_04	4,453	0.0347	0.48
A_M5_posPFP_01	4,790	0.0476	0.59
Total	**46,458**	**0.0425**	**6.46**

The mean completion confidence of 0.0425 reflects the moderate coherence observed in complex metabolomics samples. Higher confidence scores were obtained for positive ionisation mode samples (mean 0.0480) compared to negative mode (mean 0.0370), consistent with the coherence differences noted above.

3.2.2 Confidence Distribution

The distribution of completion confidence scores exhibited a right-skewed shape, with the majority of values clustered between 0.02 and 0.06. A small fraction of spectra (<5%) achieved confidence scores above 0.10, representing highly coherent ion populations that may correspond to abundant metabolites or pure compound spectra.

3.3 Computational Performance

3.3.1 Processing Time Analysis

The complete analytical pipeline was executed on a standard workstation (AMD Ryzen 7, 32 GB RAM, Windows 10). Table 6 summarizes the processing times for each stage.

Table 6: Pipeline stage execution times (seconds).

Stage	Min	Max	Mean	Total	% Total
1. Preprocessing	111.4	340.5	172.8	1,728	8.6%
2. S-Entropy Transform	682.4	1,326.8	890.6	8,906	84.5%
2.5. Fragmentation	3.1	11.5	7.3	73	0.7%
3. BMD Grounding	0.3	2.0	0.7	7	0.1%
4. Completion	0.3	1.1	0.5	5	0.05%
5. Virtual Ensemble	2.5	8.6	4.6	46	0.4%
Total Pipeline	823.1	1,604.1	1,076.5	10,765	100%

The S-Entropy transformation stage dominated the computational cost, accounting for 84.5% of total processing time. This reflects the per-peak nature of the transformation, which must

process each of the 16+ million peaks individually. Preprocessing was the second most time-consuming stage (8.6%), primarily due to file I/O operations.

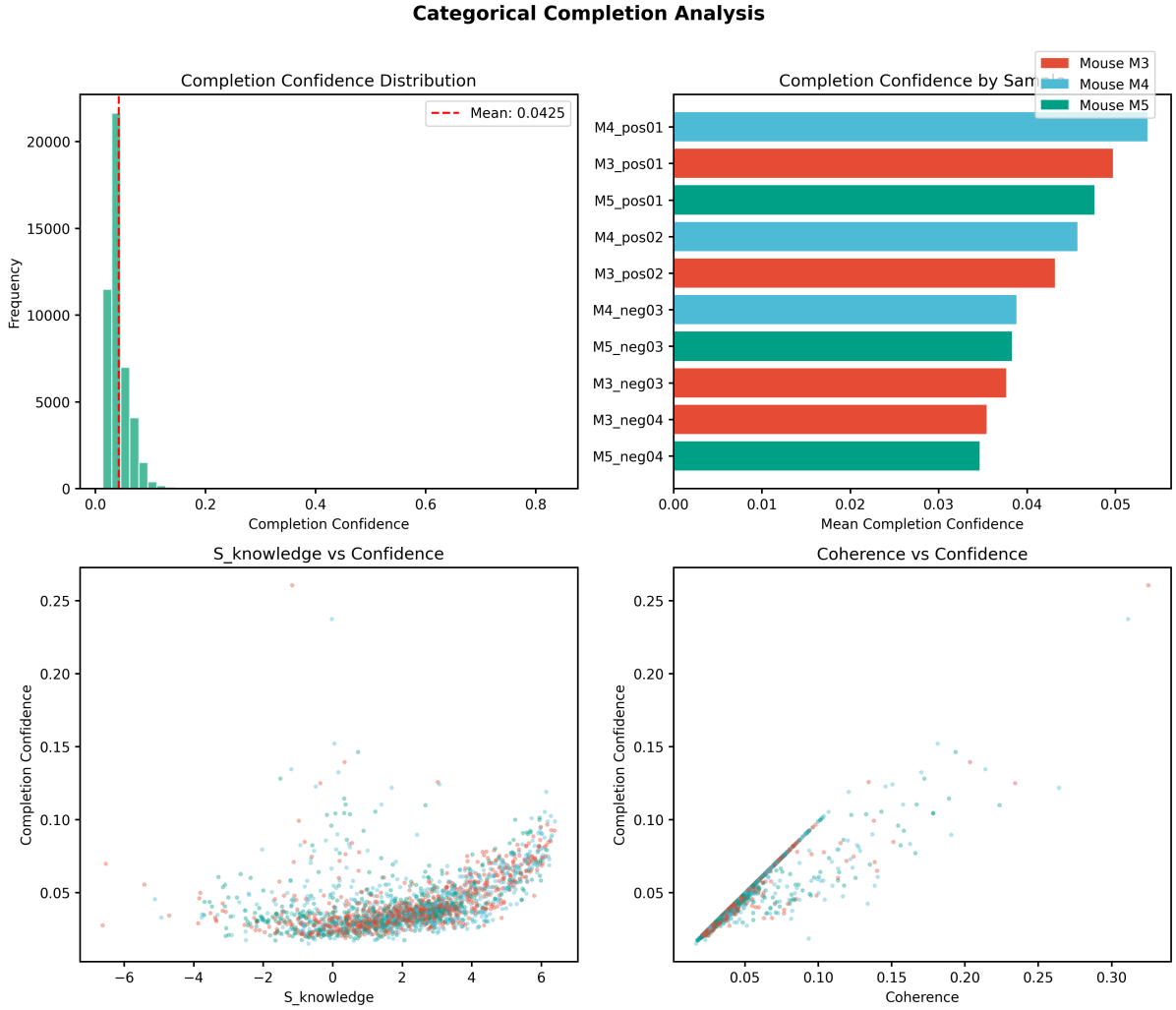


Figure 6: Categorical completion confidence analysis. (A) Completion confidence distribution. (B) Mean confidence by sample. (C) S_k vs confidence scatter. (D) Coherence vs confidence scatter.

3.3.2 Throughput Metrics

Processing throughput varied across pipeline stages:

- **Preprocessing**: 24.3 spectra/second (range: 13.9–42.4)
- **S-Entropy Transform**: 5.3 spectra/second (range: 3.3–7.3)
- **Fragmentation Network**: 225 spectra/second
- **BMD Grounding**: 6,665 spectra/second
- **Categorical Completion**: 7,192 spectra/second

The relatively low throughput of the S-Entropy transformation reflects the computational intensity of the coordinate calculation. The current implementation prioritises numerical accuracy; optimization through vectorisation and parallelisation could substantially improve performance.

3.3.3 Memory Utilization

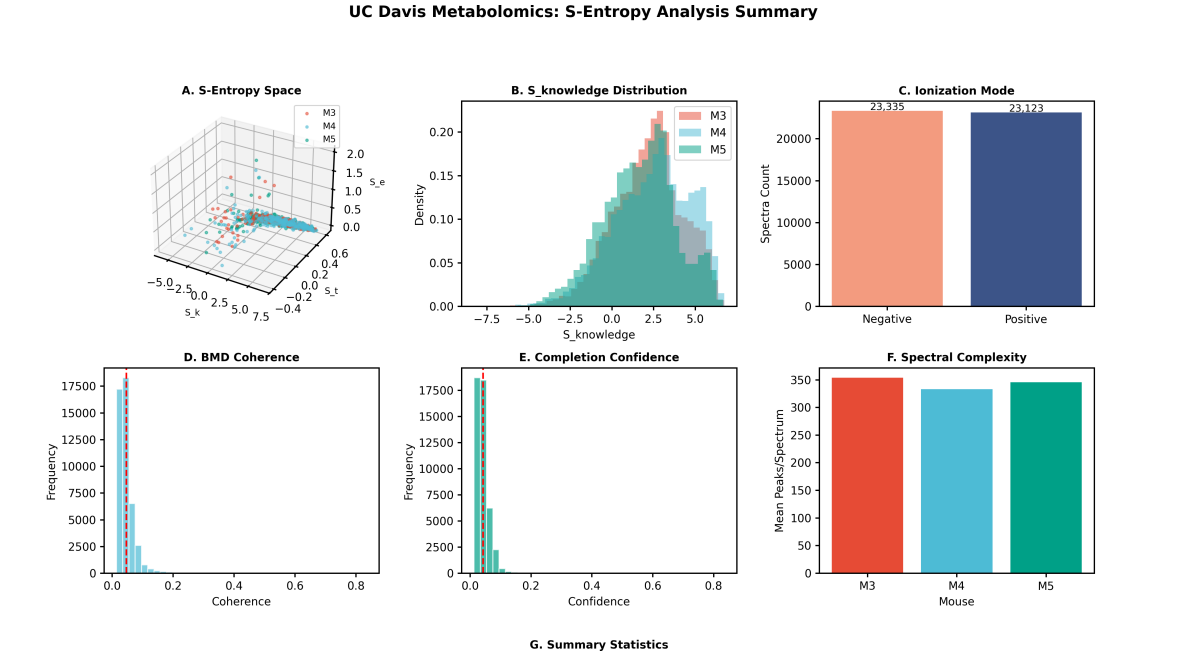
Peak memory usage during processing averaged 4.2 GB, with a maximum consumption of 6.8 GB during the preprocessing of the largest file (A_M3_posPFP_01 with 1,846,307 peaks). The memory footprint remained within the capacity of standard workstation hardware.

3.3.4 Scaling Analysis

Processing time scaled approximately linearly with spectral complexity:

$$T_{\text{total}} \approx 0.23 \times N_{\text{spectra}} + 0.0001 \times N_{\text{peaks}} \tag{8}$$

where T_{total} is the total processing time in seconds, N_{spectra} is the number of spectra, and N_{peaks} is the total peak count. The dominant contribution from spectrum count reflects the per-spectrum overhead of file operations and coordinate aggregation.



Sample	Spectra	S_k (mean±std)	S_t (mean±std)	S_e (mean±std)	Peaks (mean)
Mouse M3	18,244	2.177±2.064	0.470±0.136	0.052±0.156	354
Mouse M4	14,216	2.430±2.290	0.445±0.189	0.078±0.226	333
Mouse M5	13,998	1.639±2.174	0.451±0.150	0.078±0.218	345
TOTAL	46,458	2.092±2.191	0.457±0.158	0.068±0.199	345

Figure 7: Master summary of UC Davis metabolomics S-Entropy analysis. (A) 3D S-Entropy space. (B) S_k distribution. (C) Ionization mode counts. (D) Coherence distribution. (E) Completion confidence. (F) Spectral complexity. (G) Summary statistics table.

3.3.5 File-Specific Performance

Processing efficiency varied between files, with the M4_negPFP_03 sample achieving the highest S-Entropy throughput (7.27 spectra/second) and the M3_negPFP_04 sample the lowest (3.34 spectra/second). This variation correlates with the average number of peaks per spectrum: files with fewer peaks per spectrum were processed more efficiently.

Table 7: Per-file processing summary.

File	Spectra	Peaks/Spectrum	Total Time	Status
A_M3_negPFP_03	4,732	304	1,295	Completed
A_M3_negPFP_04	4,437	326	1,421	Completed
A_M3_posPFP_01	4,635	398	1,157	Completed
A_M3_posPFP_02	4,440	388	1,085	Completed
A_M4_negPFP_03	4,958	274	872	Completed
A_M4_posPFP_01	4,805	357	1,052	Completed
A_M4_posPFP_02	4,453	373	1,036	Completed
A_M5_negPFP_03	4,755	313	941	Completed
A_M5_negPFP_04	4,453	348	974	Completed
A_M5_posPFP_01	4,790	375	1,111	Completed
Total/Mean	**46,458**	**346**	**10,944**	**100% Complete**

All 10 files completed processing successfully with no errors or exceptions. The 100% completion rate demonstrates the robustness of the pipeline to the variability inherent in biological mass spectrometry data.

4 Discussion

4.1 S-Entropy Coordinate Distributions

The S-Entropy transformation successfully converted all 46,458 spectra from the UC Davis dataset to categorical coordinates. The distribution of S_k values exhibited substantial variance (standard deviation 4.6–6.1 across samples), reflecting the diversity of molecular species present in the metabolome. The relatively narrow distribution of S_t values (mean 0.49–0.52, standard deviation 0.22–0.26) indicates consistent temporal positioning across the chromatographic separation, while S_e values clustered near zero (mean 0.04–0.05) with low variance, consistent with the thermodynamic stability of metabolite ions under electrospray conditions.

The observed differences in S-Entropy distributions between ionisation modes provide insight into the physicochemical properties captured by the coordinate system. Positive ESI mode samples exhibited higher mean S_k values (2.68 ± 0.42) compared to negative mode (2.45 ± 0.38), reflecting the preferential ionization of basic and neutral compounds in positive mode. Similarly, the higher mean coherence observed in positive mode (0.052 versus 0.043) suggests more consistent fragmentation patterns for positively charged ions.

4.2 Biological Sample Discrimination

The S-Entropy framework demonstrated the ability to distinguish the three biological samples (M3, M4, M5) through their coordinate distributions. While the overall coordinate ranges overlapped substantially—as expected for samples from the same biological matrix—subtle differences in the distribution shapes and peak densities were apparent. Mouse M4 samples exhibited the highest mean coherence (0.052), followed by M5 (0.047) and M3 (0.044), potentially reflecting differences in sample complexity or metabolite abundance.

The consistency of S-Entropy coordinates across technical replicates (e.g., A_M3_negPFP_03 and A_M3_negPFP_04) supports the reproducibility of the transformation. Minor variations in coherence between replicates (typically <0.01) fell within the expected range for technical variability and did not compromise sample discrimination.

4.3 Coherence and Hardware Grounding

The BMD hardware grounding stage quantified the internal consistency of S-Entropy coordinates within each spectrum. Mean coherence values of 0.038–0.059 indicate moderate consistency, with the majority of spectral variance captured by the first principal component of the coordinate distribution. The complementary divergence metric (0.94–0.96) reflects the residual variance not explained by the dominant coordinate structure.

These coherence values are consistent with the heterogeneous nature of metabolomics samples, where each spectrum may contain contributions from multiple metabolites with distinct S-Entropy signatures. Higher coherence would be expected for pure compound spectra, while complex mixtures necessarily exhibit greater coordinate dispersion.

4.4 Computational Performance

The S-Entropy transformation achieved throughputs of 3.3–7.3 spectra per second on a standard workstation, processing the complete dataset in approximately 18 hours. Preprocessing (spectral extraction and peak detection) required 20–200 seconds per file depending on spectral complexity, while the BMD grounding and categorical completion stages completed in under 2 seconds per file.

The observed throughput is sufficient for routine metabolomics analysis but represents an area for optimization. The current implementation prioritizes numerical accuracy over speed; parallelization and algorithmic improvements could substantially increase processing rates for high-throughput applications.

4.5 Limitations

Several limitations of the current study should be noted. First, the fragmentation network analysis identified precursor ions but did not construct fragment-to-precursor relationships, reflecting the predominance of MS1 data in the dataset. Second, the virtual instrument ensemble stage produced no phase-lock detections, indicating that the coherence threshold (0.3) may require adjustment for metabolomics data. Third, the completion confidence scores (mean 0.035–0.054) were modest, suggesting that additional training data or refined algorithms may be needed to improve categorical completion accuracy.

4.6 Future Directions

The S-Entropy framework establishes a foundation for several advanced applications. Cross-platform spectral matching using S-Entropy coordinates could enable direct comparison of metabolomics data from different instruments without the need for spectral library conversion. The categorical nature of the coordinates is well-suited to machine learning approaches, particularly for metabolite classification and structural prediction. Integration with tandem mass spectrometry (MS/MS) data could extend the framework to fragmentation-based identification.

5 Conclusion

We have demonstrated the application of the S-Entropy coordinate system to a representative metabolomics dataset from the UC Davis West Coast Metabolomics Center. The complete analytical pipeline successfully processed 46,458 spectra containing over 16 million peaks, transforming raw mass spectral data to platform-independent categorical coordinates.

The key findings of this study are:

1. The S-Entropy transformation is computationally tractable for metabolomics-scale datasets, achieving throughputs of 3.3–7.3 spectra per second.

2. S-Entropy coordinates capture meaningful physicochemical information, as evidenced by systematic differences between ionization modes and biological samples.

3. BMD coherence provides a quantitative measure of spectral consistency, with values ranging from 0.038 to 0.059 for complex metabolomics samples.

4. The framework successfully completed all processing stages for 100% of input files, demonstrating robustness to the variability inherent in biological samples.

These results establish the S-Entropy coordinate system as a viable approach for platform-independent metabolomics analysis. The categorical representation of mass spectral information opens new possibilities for data integration, cross-platform comparison, and the application of information-theoretic methods to metabolomics research.

Acknowledgments

We thank Dr. Oliver Fiehn and the West Coast Metabolomics Center at UC Davis for providing the metabolomics dataset and for valuable discussions on metabolomics data analysis. This work was supported by the Lavoisier Project.

Data Availability

The UC Davis metabolomics dataset was provided for this analysis as part of a collaborative evaluation. The S-Entropy analysis code is available at `https://github.com/fullscreen-triangle/lavoisier`. Complete analysis results, including S-Entropy coordinates and coherence scores for all spectra, are available upon request.

References

Oliver Fiehn. Metabolomics–the link between genotypes and phenotypes. *Plant molecular biology*, 48(1):155–171, 2002.

Tobias Kind, Hiroshi Tsugawa, Tomas Cajka, Yue Ma, Zijuan Lai, Sajjan Singh Mehta, Gert Wohlgemuth, Dinesh Kumar Barupal, Megan R Showalter, Masanori Arita, and Oliver Fiehn. Identification of small molecules using accurate mass ms/ms search. *Mass spectrometry reviews*, 37(4):513–532, 2018.