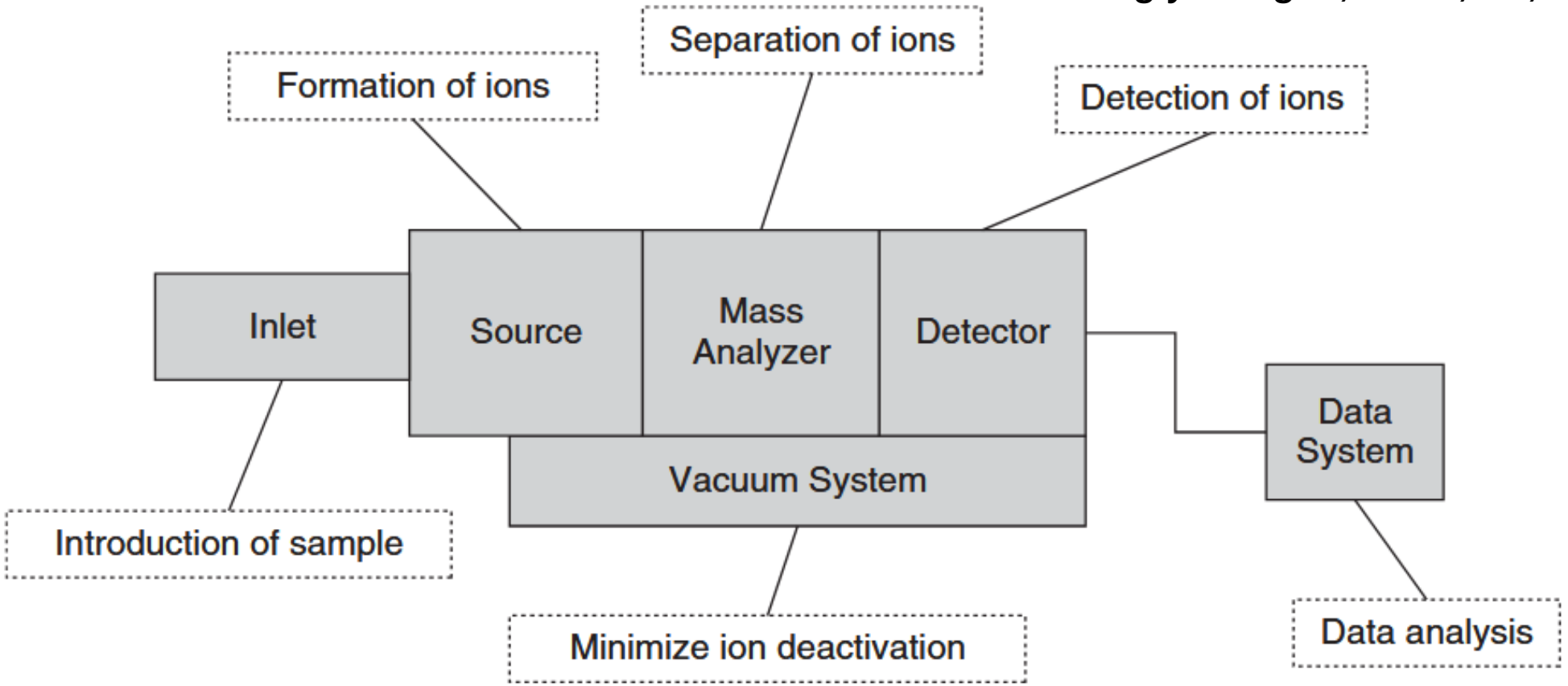
Probabilistic Persistent Homology for Peak Extraction in Archive-Scale Mass Spectrometry
概率持续同调在归档级质谱分析中的峰提取应用

Shao Shi (石邵), PhD. Candidate
School of Environment & CS, SUSTech
2026.5.21

Mass Spectrometry (MS，质谱)

m/z, mass-to-charge ratio, 质荷比
If singly charged, z = 1, i.e., m



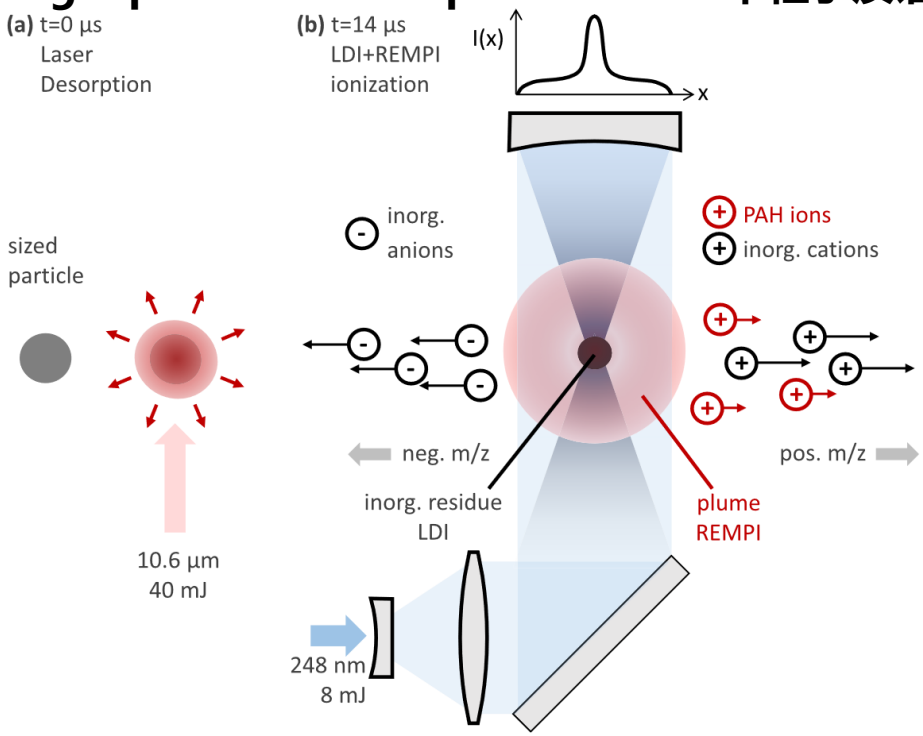
Traditional MS involves prior separation of chemicals (e.g., by Liquid Chromatography(LC) / Gas Chromatography (GC)) before the inlet of mass spectrometer.

Online MS 在线质谱

Techniques like SPMS, PTR-TOF-MS, AMS/ACSM, and CIMS capture thousands of species in real-time (e.g., transient VOCs, secondary organic aerosols) with exceptionally high detection speed.

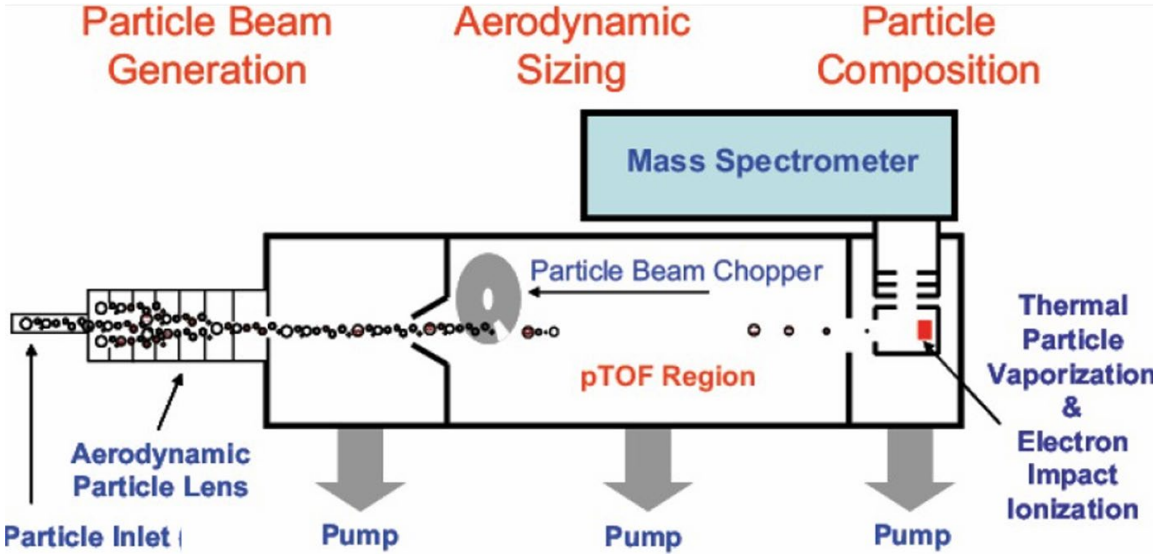
- **The Trade-off:** By eliminating prior chromatographic separation, mass spectra become highly congested and temporally dynamic.
- **The Bottleneck:** Extracting accurate, *untargeted* chemical information from dense, multidimensional datasets without manual bias.

Single particle mass spectrometer 单粒子质谱仪



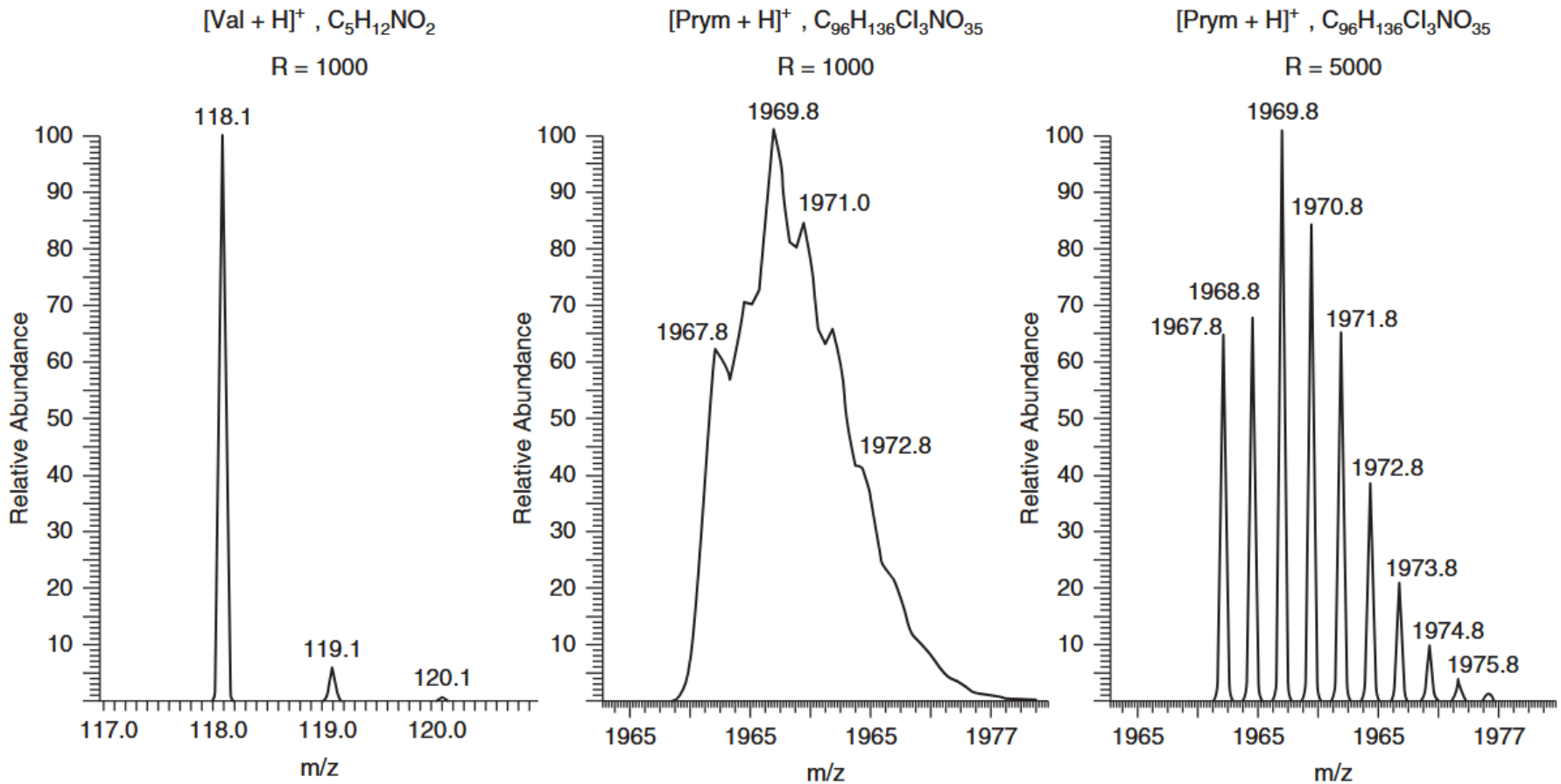
Schade, J. *et al. Analytical Chemistry.* 91, 10282–10288 (2019).

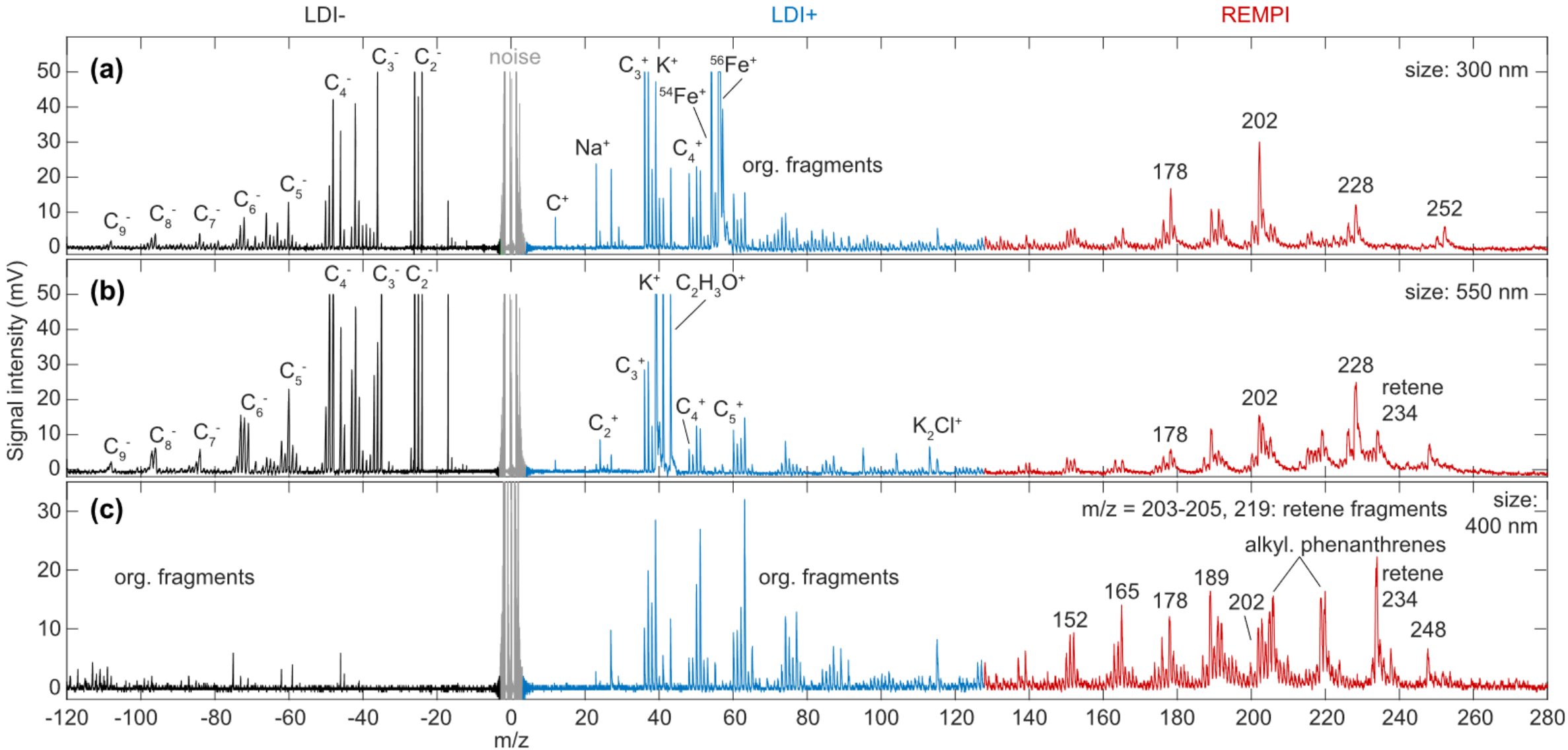
Aerosol Mass Spectrometer 气溶胶质谱仪



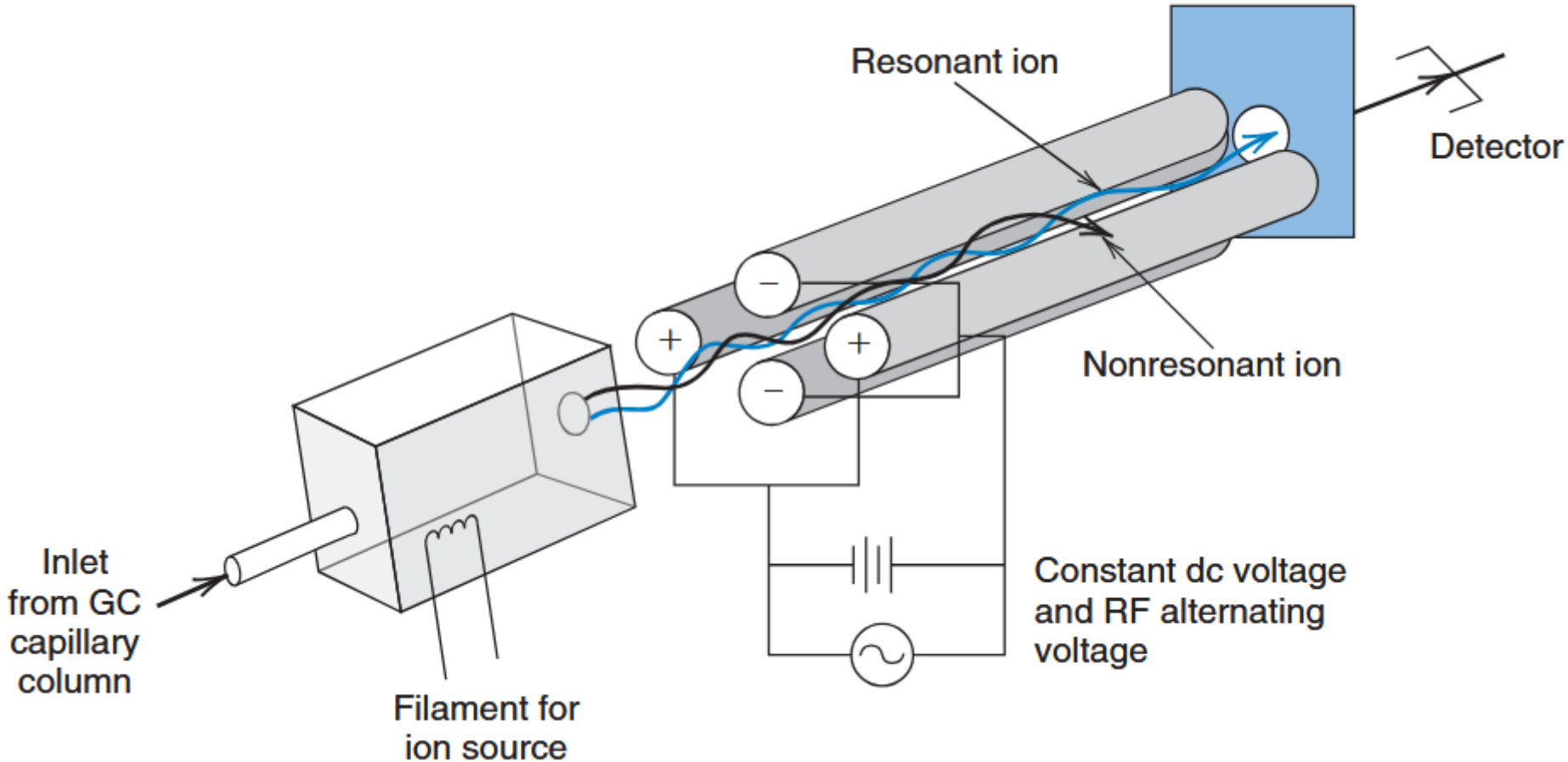
https://aerodyne.com

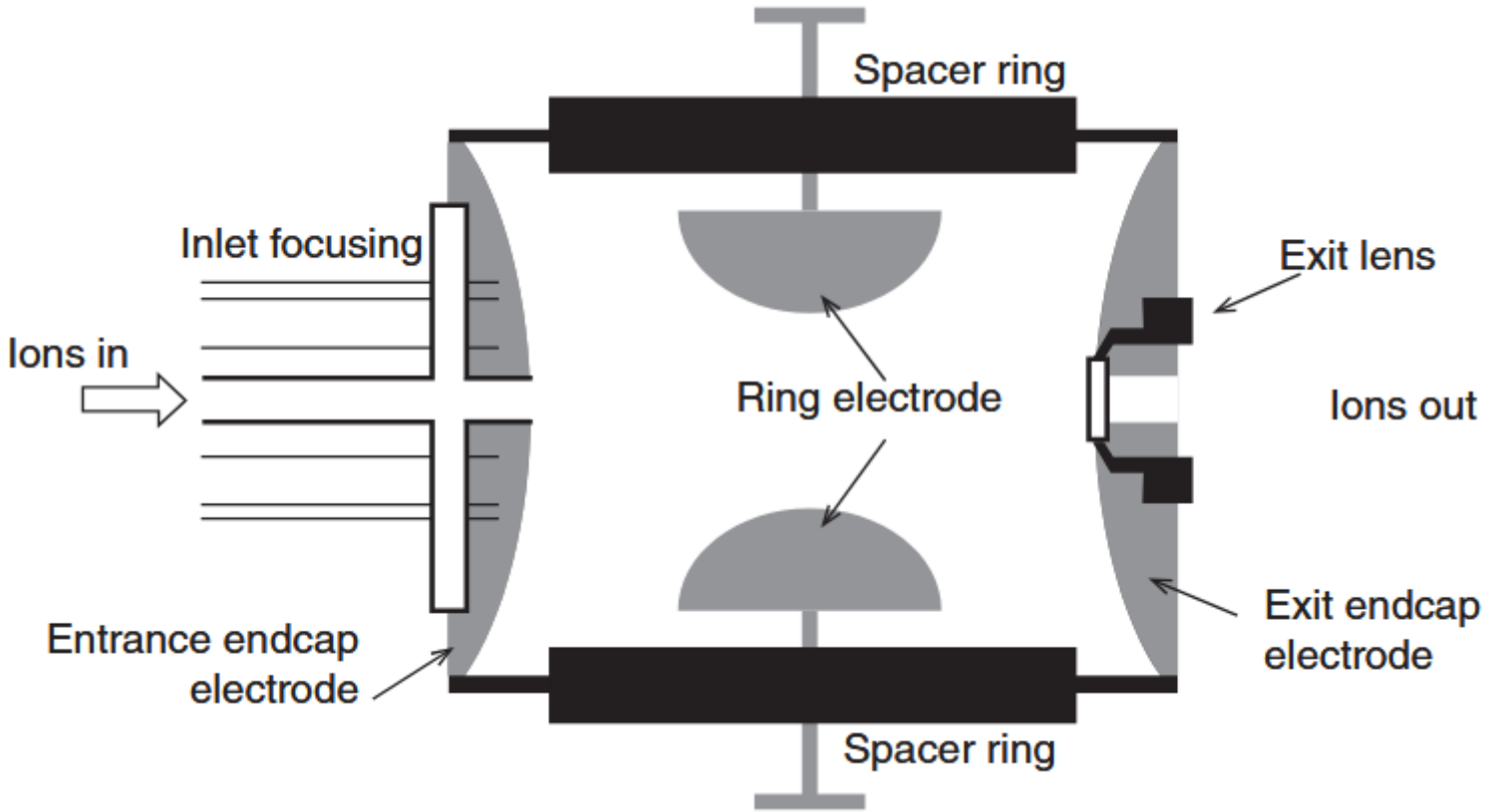
$$R = \frac{m}{\Delta m}$$

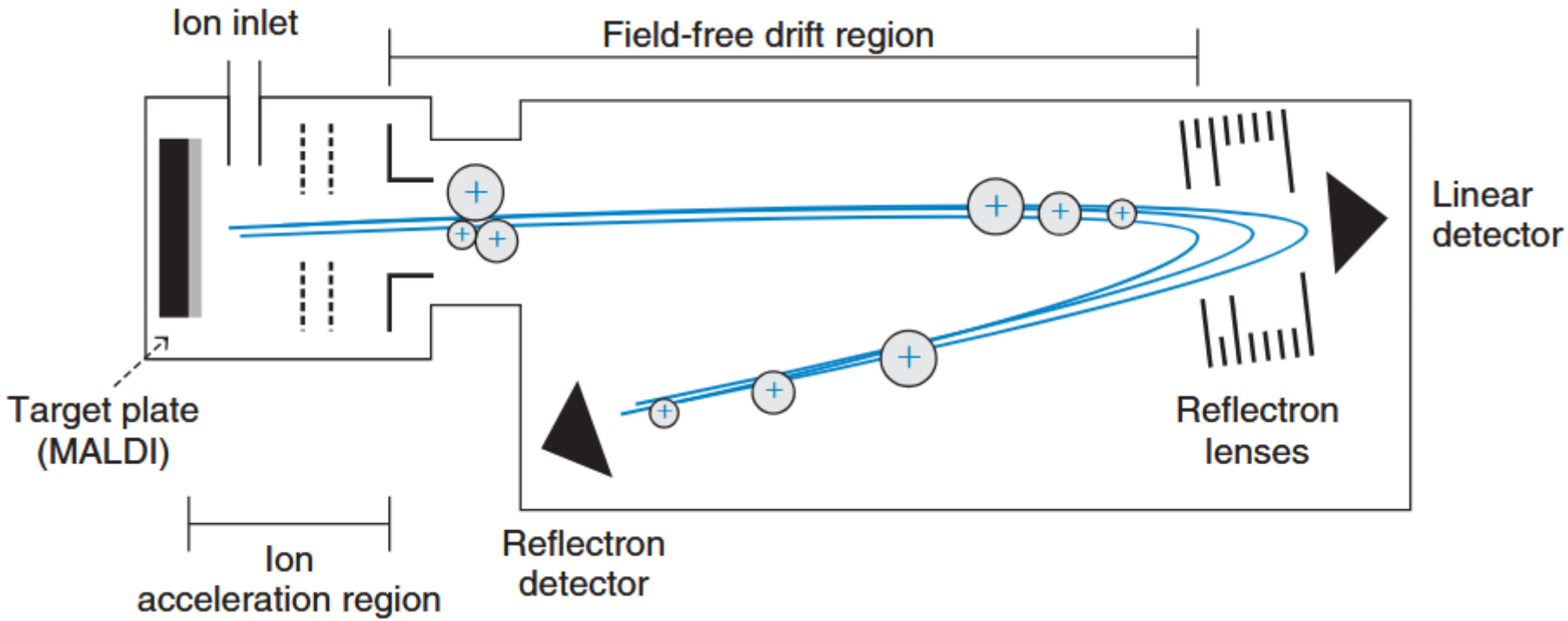


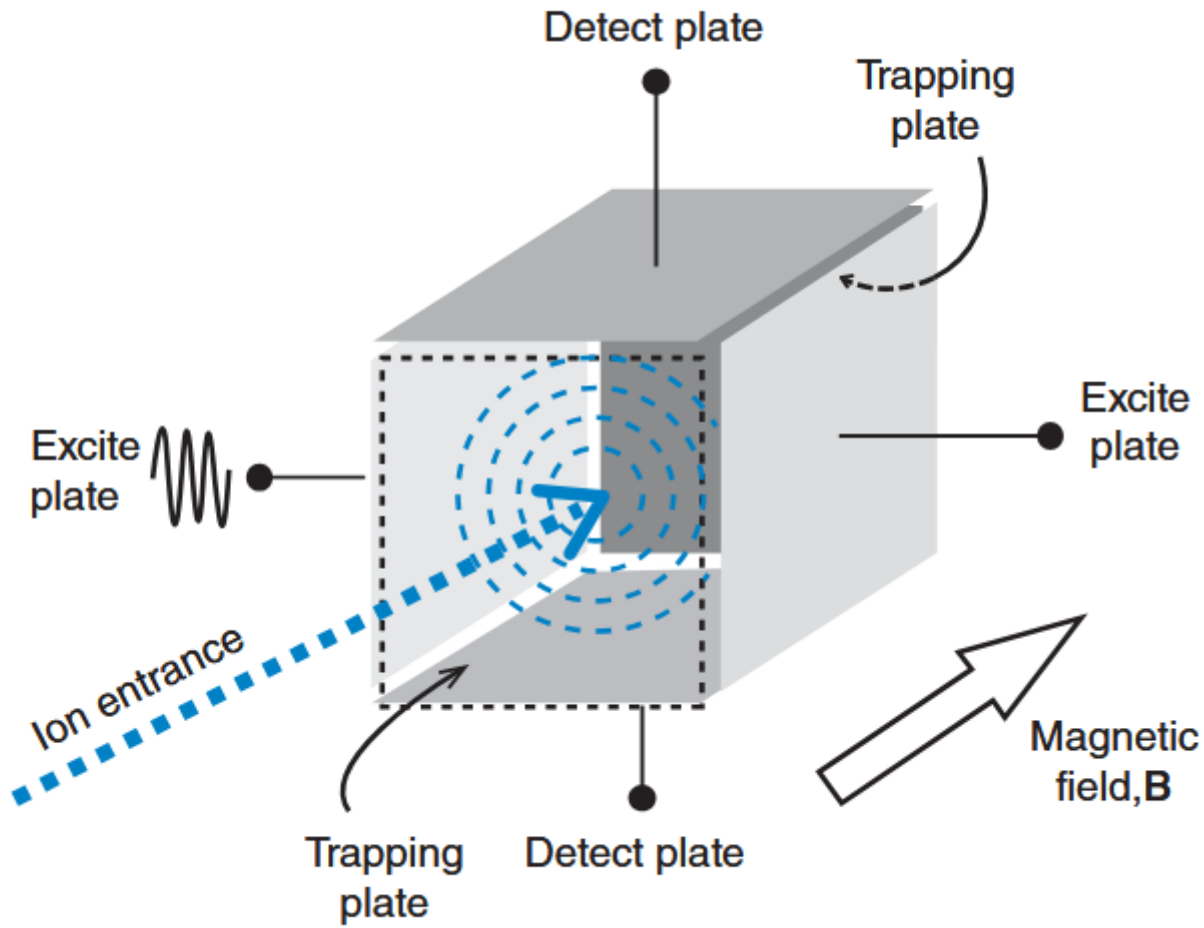


baseline drift(基线漂移)


Resonant ion
Detector
Nonresonant ion
Inlet
from GC
capillary
column
Filament for
ion source
Constant dc voltage
and RF alternating
voltage

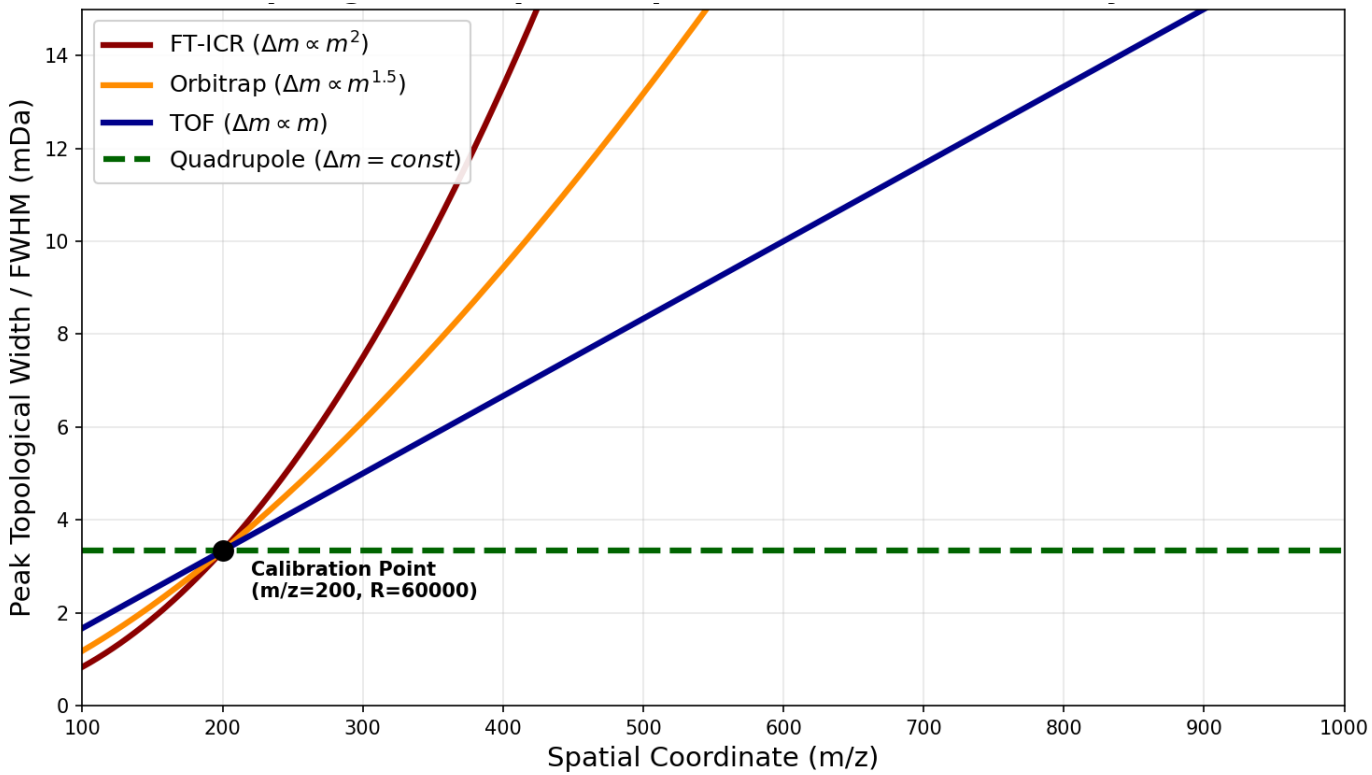

Spacer ring
Inlet focusing
Ions in
Exit lens
Ring electrode
Ions out
Entrance endcap
electrode
Exit endcap
electrode
Spacer ring


Ion inlet
Field-free drift region
Linear
detector
Target plate
(MALDI)
Reflectron
lenses
Ion
acceleration region
Reflectron
detector


Detect plate
Trapping plate
Excite plate
Excite plate
Ion entrance
Trapping plate
Detect plate
Magnetic field, B

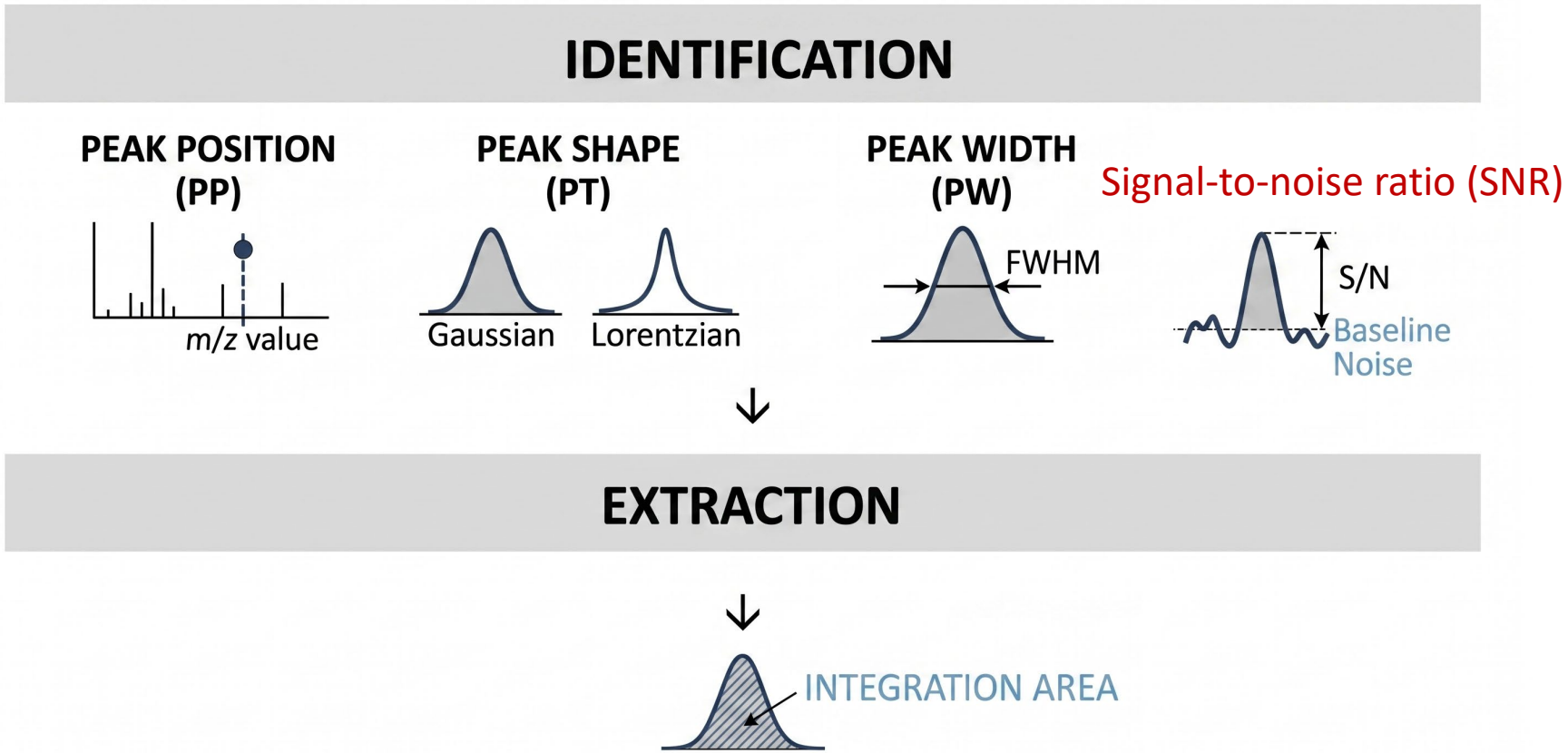
- **The Archive-Scale Problem:** Big data MS repositories contain data generated by fundamentally different physical hardware.
- **Topology is not static:** The spatial footprint of a peak (FWHM or Δm) diverges rapidly across the spatial coordinate (m/z):
  - **Quadrupole:** Constant Width ($\Delta m = \text{const}$)
  - **TOF:** Linear Growth ($\Delta m \propto m$)
  - **Orbitrap:** Exponential Growth ($\Delta m \propto m^{1.5}$)
  - **FT-ICR:** Quadratic Growth ($\Delta m \propto m^2$)

A fixed search window fails.



The Data Deluge: Modern online MS produce massive datasets (>10 Hz), demanding real-time "edge computing" processing.

Our Goal: Develop a deterministic, non-iterative algorithm that guarantees computational speed, physical accuracy, and insensitive to noise and baseline drift.



Current Approach for Peak Extraction in Online MS

Top-Down (Model-Driven)

Examples: **Tofware, PIKA, multi-peak fitting**

How it works: Imposes idealized mathematical shapes (e.g., Gaussian) onto the data.

The Good: Unmatched at quantifying *known* overlapping isobars when guided by a **library**.

The Bad (for Untargeted Analysis):

- Relies on "residual discovery" to find unknowns.
- Accurate mass calibration required.
- **Rigid:** Misinterprets natural peak tailing as hidden, fake peaks (High false positives).
- **Fragile:** Minor mass calibration shifts break the model (Local-minima trapping).
- **Expensive.**

Bottom-Up (Data-Driven)

Examples: **Derivatives, Standard CWT**

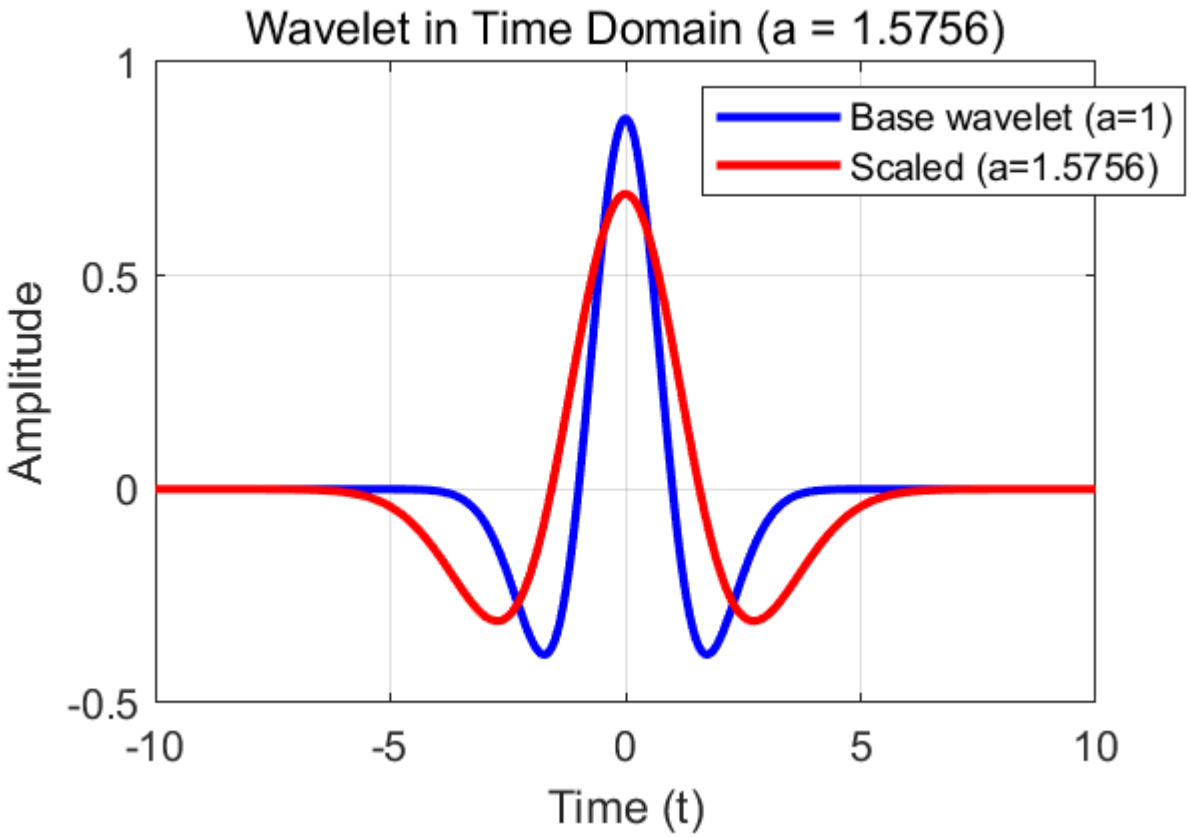
How it works: Extracts features based on local signal patterns without rigid shape rules.

The Good:

- **Library-Free:** Perfect for blind, exploratory discovery.
- **Robust:** Naturally filters baseline drift and ignores mass calibration errors (mass calibration free).

The Challenge: Standard methods remains mathematical, lack physical limits, often over-extracting noise on peak shoulders.

CWT



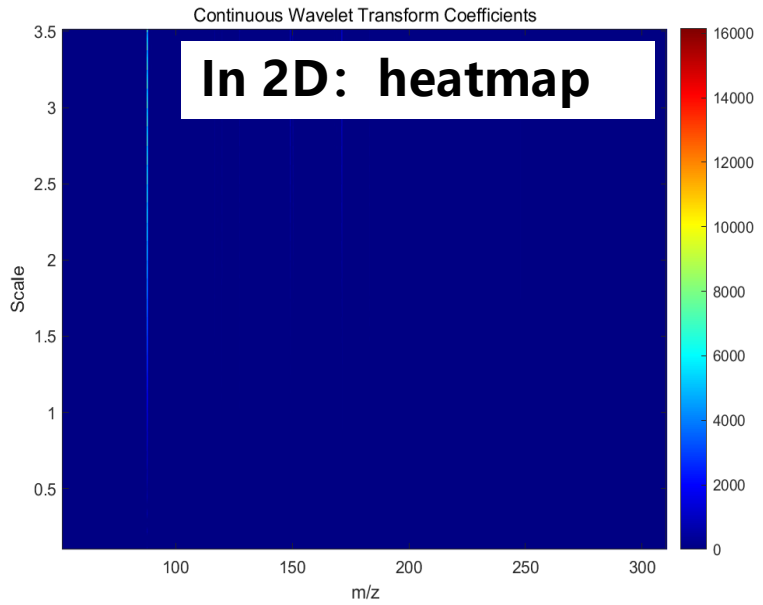
Ricker wavelet

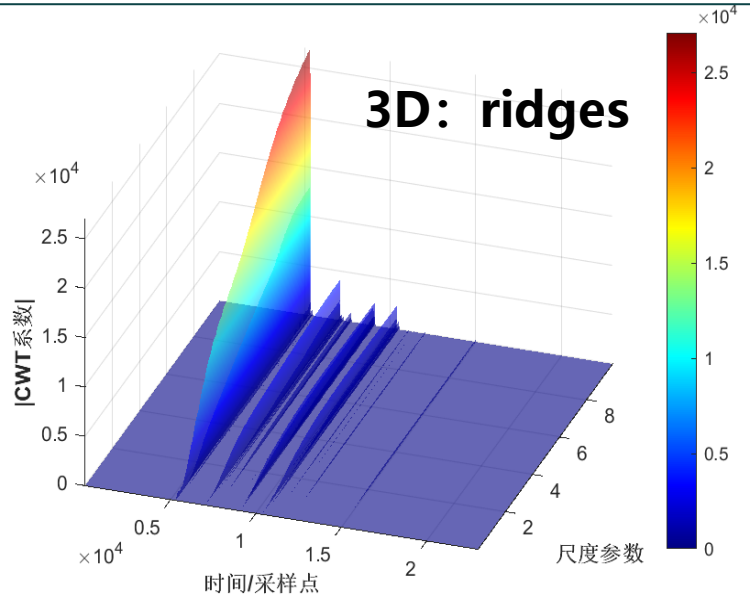
$$\psi_a(t) = \frac{1}{\sqrt{a}}\psi\left(\frac{t}{a}\right) = \frac{1}{\sqrt{a}}\frac{1}{\sigma_w^2}\left(1 - \frac{t^2}{a^2\sigma_w^2}\right)e^{-\frac{t^2}{2a^2\sigma_w^2}}$$

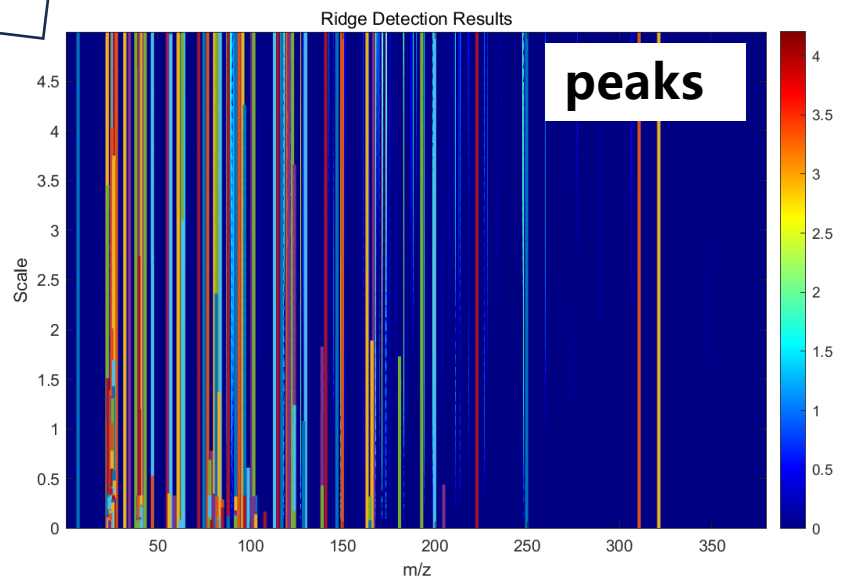
$$W(s,b) = \int_{-\infty}^{\infty} f(t)\frac{1}{\sqrt{s}}\psi\left(\frac{t-b}{s}\right)dt$$

It can be proven (proof omitted here) that the Mexican hat wavelet at scale yields the maximum response to a Gaussian peak with $\sigma_g = \frac{a}{\sqrt{2}}$.

CWT







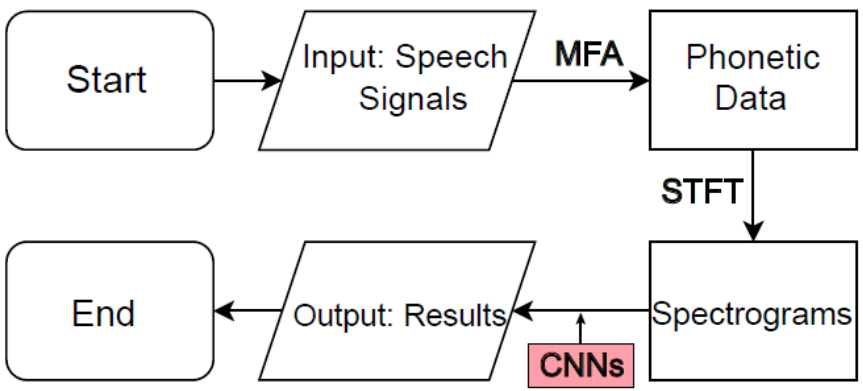
- The ridge lines in the CWT coefficient matrix correspond one-to-one with the mass spectral peaks (ideally).
- The key to CWT-based peak identification lies in **ridge line extraction.**
- **Generally, the gradient method is used to extract ridge lines, which limits efficiency and accuracy.**

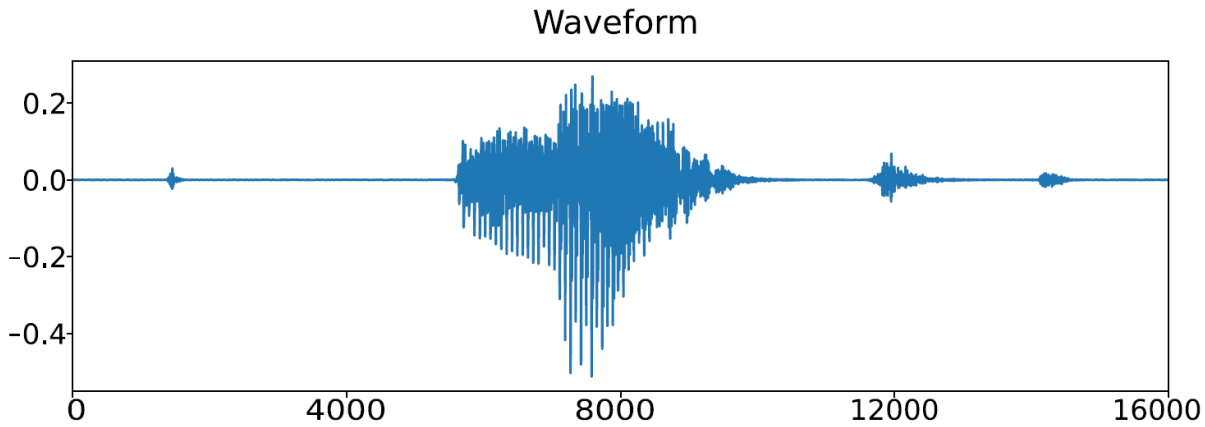
Topology-Informed Convolution Kernels for Speech Recognition Tasks

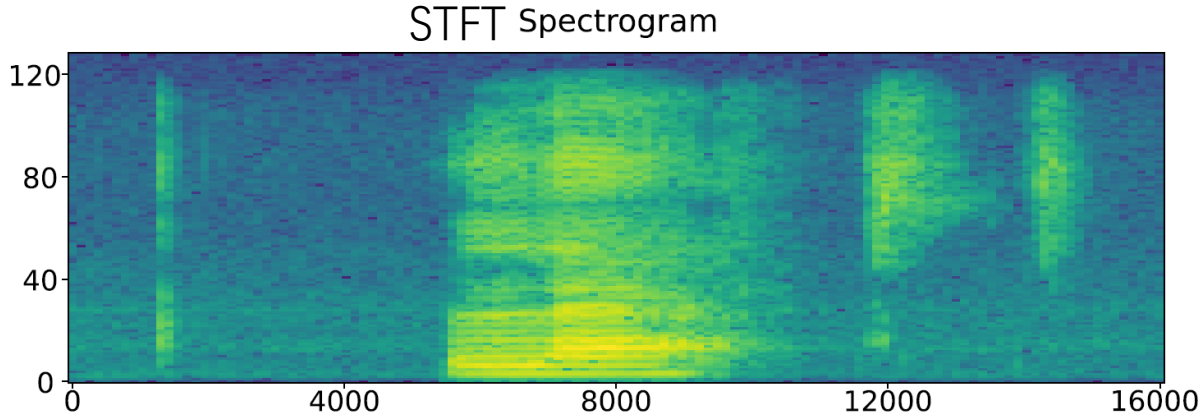
Zhiwang Yu
School of Mathematics and Systems Science
Shenyang Normal University
Shenyang, Liaoning, P.R. China

Pingyao Feng
Department of Mathematics
North Carolina State University
Raleigh, North Carolina, USA

Qingrui Qu
Haiyu Zhang
Yifei Zhu
Department of Mathematics
Southern University of Science and Technology
Shenzhen, Guangdong, P.R. China







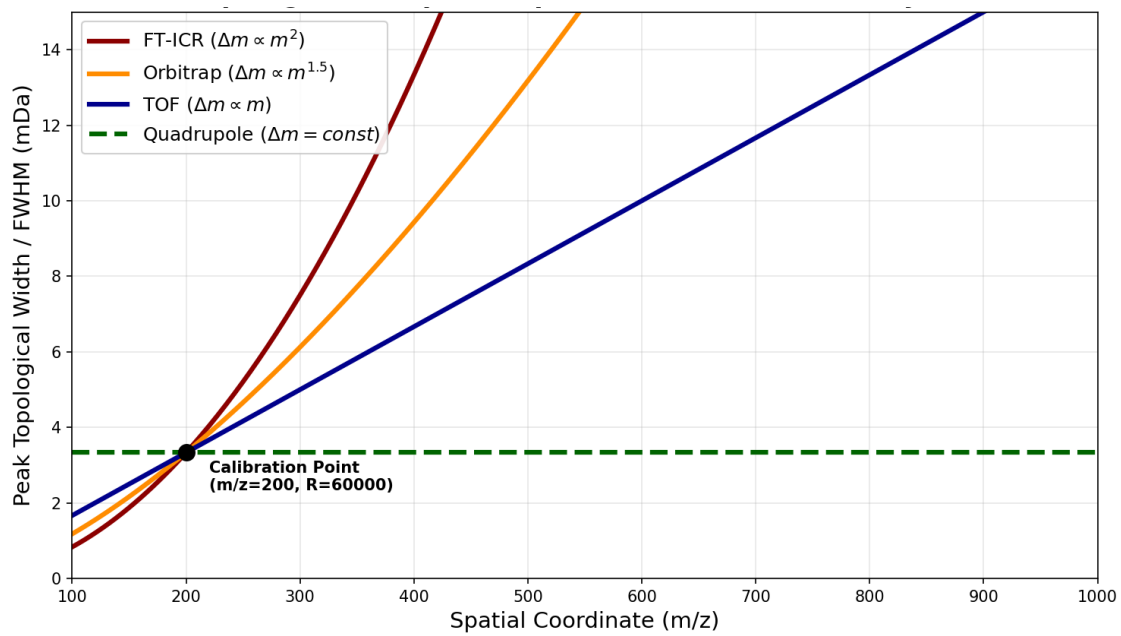
$$\mathbf{A}_1 = \mathbf{Q}\begin{bmatrix} 1 & 0 & -1 \\ 1 & 0 & -1 \\ 1 & 0 & -1 \end{bmatrix}/\sqrt{6} \text{ and } \mathbf{A}_2 = \mathbf{Q}\begin{bmatrix} 1 & -2 & 1 \\ 1 & -2 & 1 \\ 1 & -2 & 1 \end{bmatrix}/\sqrt{18}$$

Methods

Preliminaries & Definitions

Let $\mathcal{X} \subset \mathbb{R}$ be a discrete, uniformly sampled 1D domain with spatial resolution ∇x.
Let the input signal be a function $f : \mathcal{X} \to \mathbb{R}_{\geq 0}$.
Let $W : \mathcal{X} \to \mathbb{R}_{>0}$ be the theoretical Full-Width at Half-Maximum (FWHM) function.

We seek to identify a set of true structurally significant features (topological generators) $\mathcal{K} \subset \mathcal{X}$, effectively separating them from background noise and baseline drift.



Phase I: Scale-Space Construction (Continuous Wavelet Transform)

We embed $f(x)$ into a 2D scale-space manifold to compute the structural resonance of the signal across multiple resolutions.

1. The Wavelet Kernel

We define the Ricker wavelet (the negative normalized second derivative of a Gaussian) at scale $s \in \mathbb{R}_{>0}$:

$$\psi(x, s) = \frac{2}{\sqrt{3s}\pi^{1/4}} \left(1 - \frac{x^2}{s^2}\right) \exp\left(-\frac{x^2}{2s^2}\right)$$

2. The Transform & Regularization

The Continuous Wavelet Transform (CWT) of the padded signal is:

$$\mathcal{W}(x, s) = (f * \psi(\cdot, s))(x)$$

$f(x) * \psi(x) \Leftrightarrow \mathcal{F}\{f\} \cdot \mathcal{F}\{\psi\}$

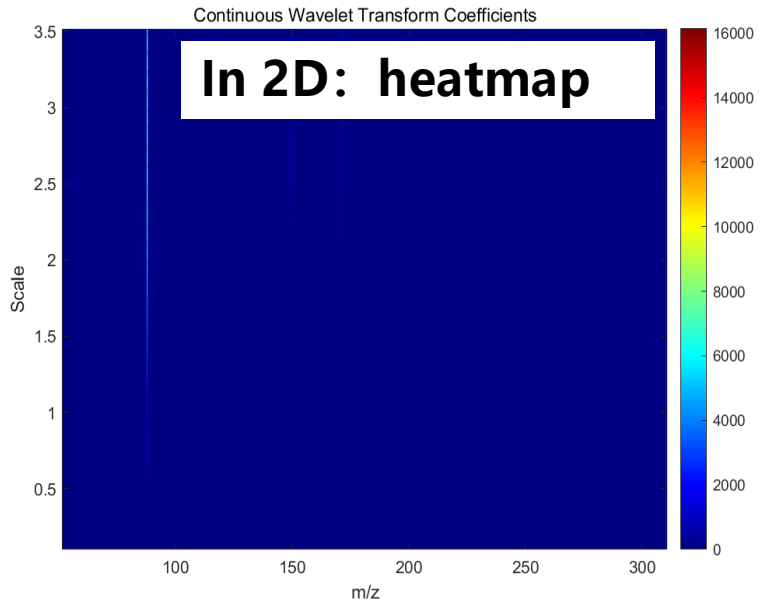
Use FFT to accelerate the computation.

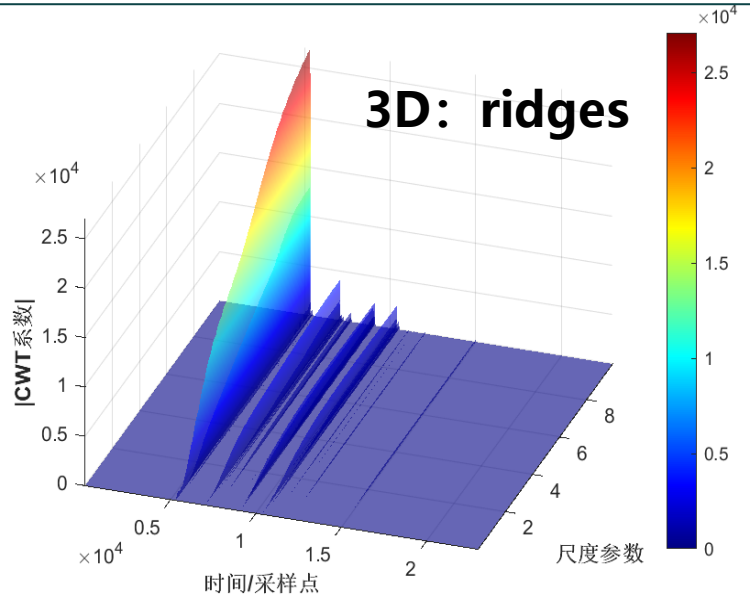
$O(N^2)$ to $O(N \log N)$

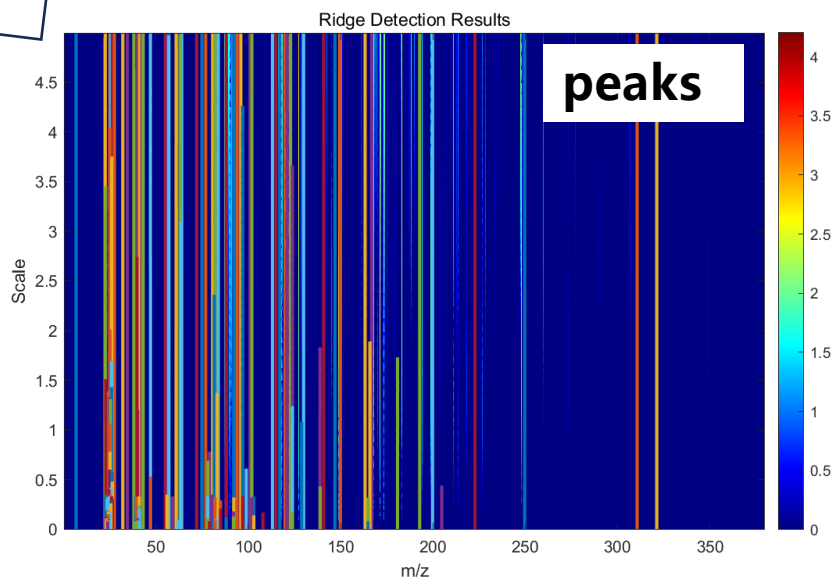
To regularize the manifold and eliminate spurious derivative artifacts (approximating a proper Morse function), we apply a scale-dependent Gaussian smoothing kernel $G_s(x) \propto \exp(-x^2/2)$:

$$\mathcal{E}(x, s) = (\mathcal{W}(\cdot, s) * G_s)(x)$$

CWT

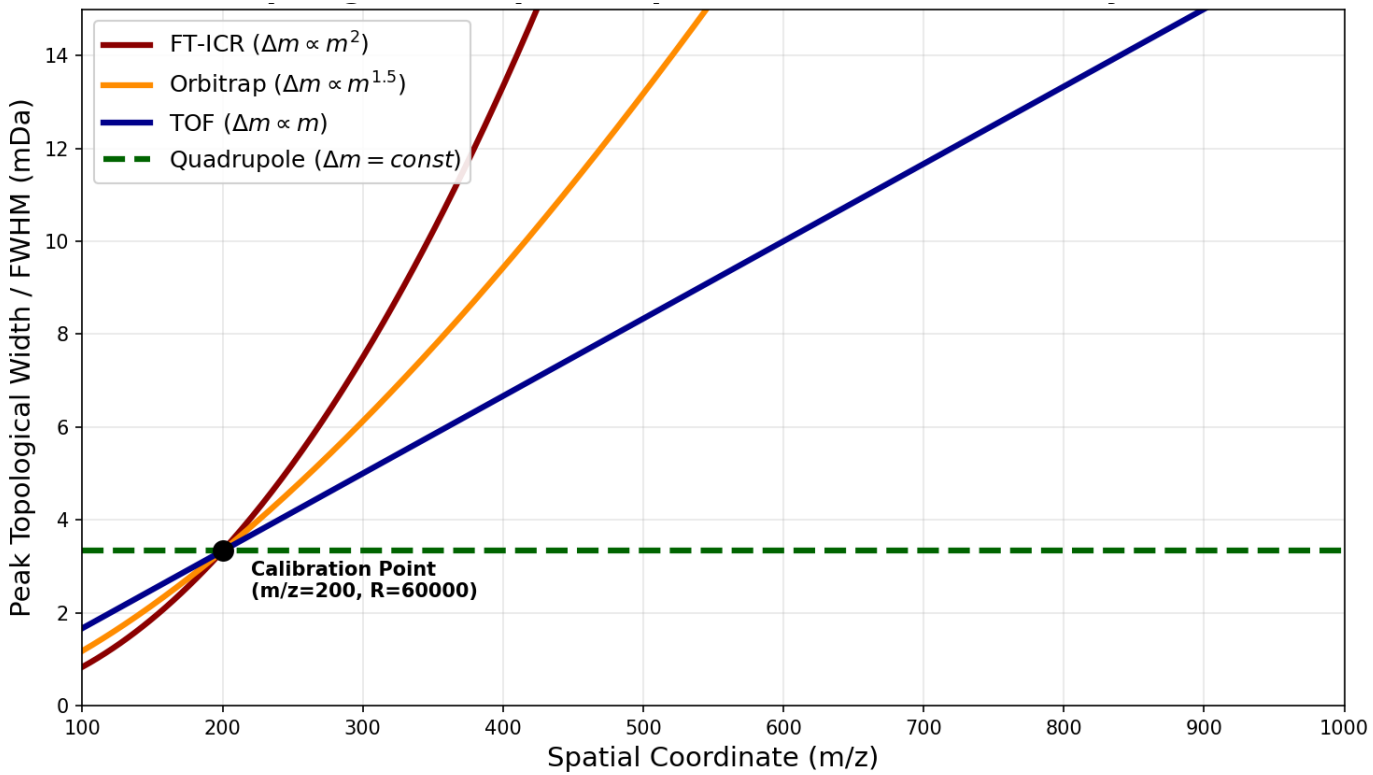






- The ridge lines in the CWT coefficient matrix correspond one-to-one with the mass spectral peaks (ideally).
- The key to CWT-based peak identification lies in **ridge line extraction.**
- **Generally, the gradient method is used to extract ridge lines, which limits efficiency and accuracy.**

- **The Archive-Scale Problem:** Big data MS repositories contain data generated by fundamentally different physical hardware.
- **Topology is not static:** The spatial footprint of a peak (FWHM or Δm) diverges rapidly across the spatial coordinate (m/z):
  - **Quadrupole:** Constant Width ($\Delta m = \text{const}$)
  - **TOF:** Linear Growth ($\Delta m \propto m$)
  - **Orbitrap:** Exponential Growth ($\Delta m \propto m^{1.5}$)
  - **FT-ICR:** Quadratic Growth ($\Delta m \propto m^2$)



The Scale-Space Solution

A fixed search window fails.

Our framework dynamically injects these theoretical FWHM trajectories into the 2D CWT scale-space as an analytical "Teardrop Constraint," forcing the topology engine to adapt to the local hardware physics.

Phase II: The "Teardrop" Attenuation Constraint

To constrain topological features to physically plausible geometries, we apply a localized analytical weighting function.

Let the optimal theoretical scale at x be defined by the physical FWHM constraint:

$$s_{opt}(x) = \frac{W(x)}{2\sqrt{2\ln 2} \cdot \nabla x} \approx \frac{W(x)}{2.355\nabla x}$$

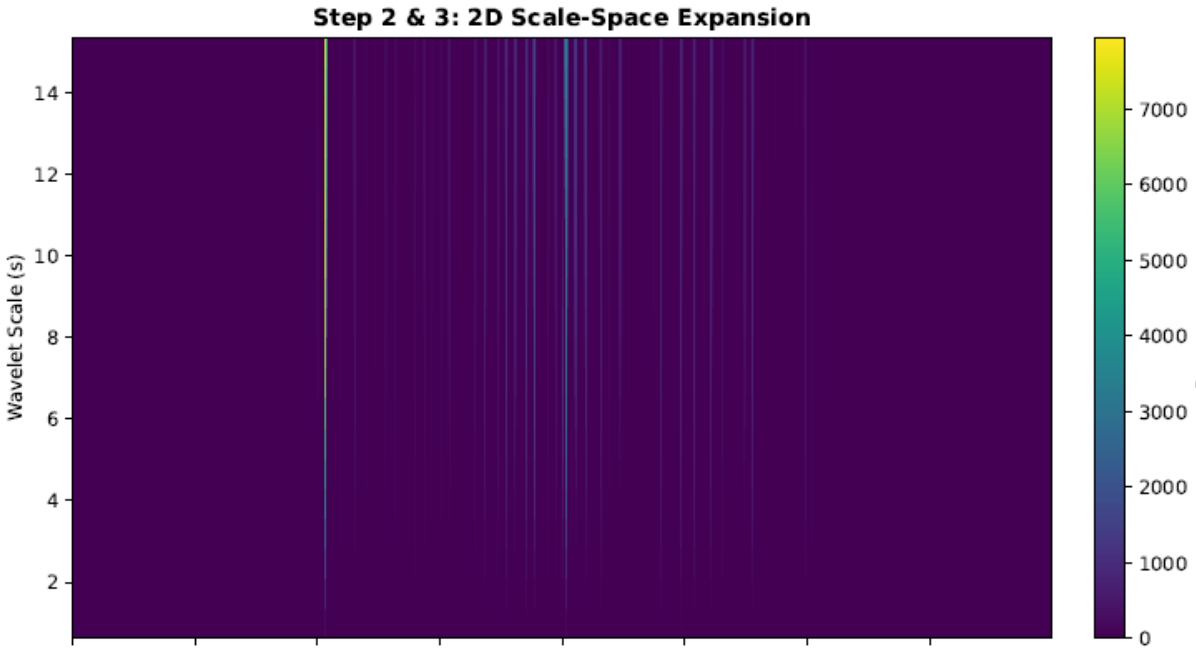
We define a spectral attenuation map $\mathcal{V} : \mathcal{X} \times \mathbb{R}_{>0} \to (0, 1]$:

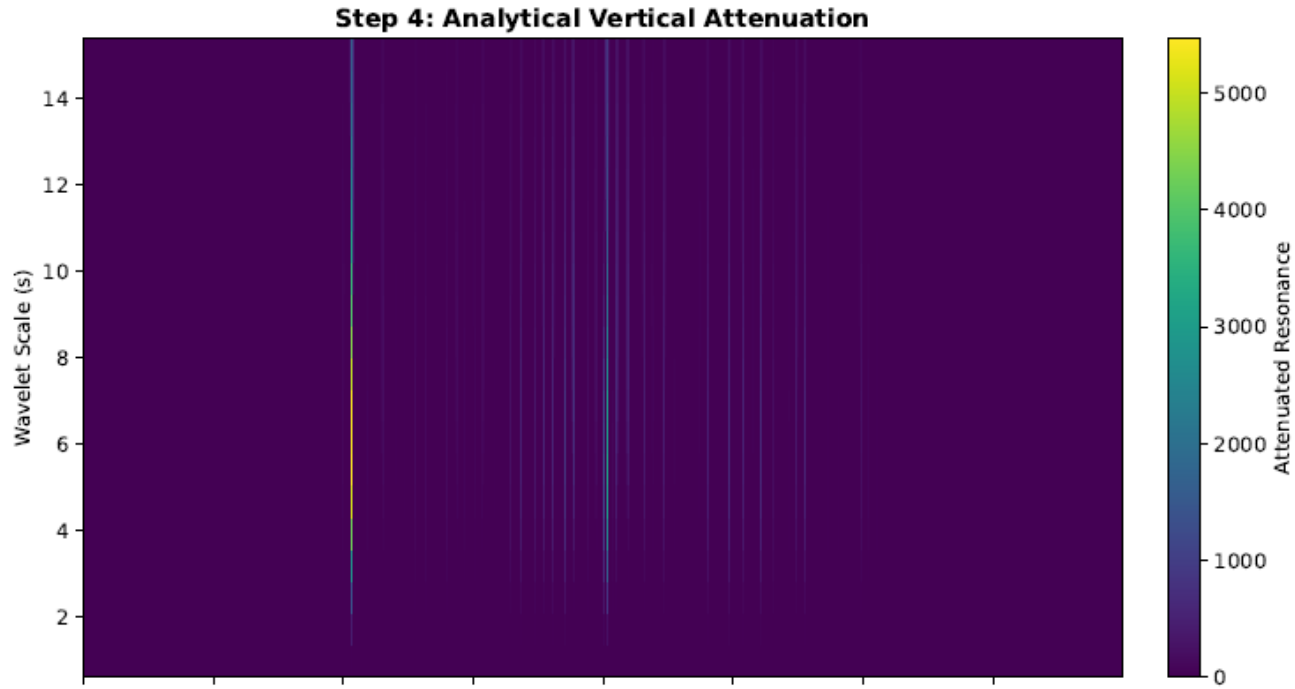
$$\mathcal{V}(x, s) = \left(\frac{2 \cdot s \cdot s_{opt}(x)}{s^2 + s_{opt}(x)^2} \right)^{5/2}$$

(Note: This function peaks at 1 when $s = s_{opt}(x)$ and decays rapidly as s deviates, forming a "teardrop" shaped boundary in scale-space).

The attenuated energy surface is:

$$\Phi(x, s) = \mathcal{E}(x, s) \cdot \mathcal{V}(x, s)$$


Step 2 & 3: 2D Scale-Space Expansion
Wavelet Scale (s)
14
12
10
8
6
4
2
7000
6000
5000
4000
3000
2000
1000
0


Step 4: Analytical Vertical Attenuation
Wavelet Scale (s)
14
12
10
8
6
4
2
5000
4000
3000
2000
1000
0
Attenuated Resonance

Phase III: Critical Point Extraction & Interpolation

We identify candidate features by extracting the non-degenerate local maxima (Index-2 critical points in the 2D scale-space) of the surface $\Phi(x, s)$ that exceed a dynamically computed local noise threshold $T(x, s)$.

Let $\mathcal{M}$ be the discrete set of candidate points:

$$\mathcal{M} = \left\{ (x, s) \in \mathcal{X} \times \mathbb{R}_{>0} \;\middle|\; \nabla\Phi(x, s) = \mathbf{0},\ \lambda_{max}(\mathbf{H}_\Phi) < 0,\ \Phi(x, s) > T(x, s) \right\}$$

For each $x_i \in \mathcal{M}$, we perform a 2nd-order Taylor expansion (parabolic interpolation) to find the sub-pixel continuous maximum x_i^*:

$$x_i^* = x_i - \frac{f'(x_i)}{f''(x_i)} \nabla x_i$$

We define the baseline-subtracted intensity function as $\tilde{f}(x) = f(x) - B(x)$, where $B(x)$ is the morphological baseline (computed via sequential infimum and mean filters).

Phase IV: Probabilistic Persistent Homology

We now evaluate the 0-dimensional Persistent Homology (H_0) of the candidates using a modified **Superlevel Set Filtration**.

1. The Elder Rule Filtration

We order the candidate points $\{x_k \in \mathcal{M}\}$ into a sequence I sorted by descending topological resonance $\Phi(x_k, s_k)$. This ordering simulates the "sea-level drop" filtration.

2. Physical Baseline Confidence

For each candidate i, we map its Signal-to-Noise Ratio (SNR) into a continuous probability measure bounded by $[0, 1)$ using an exponential saturation function governed by decay parameter β:

$$C_{phys}(i) = 1 - \exp\left(-\beta \frac{\tilde{f}(x_i)}{\sigma_{noise}}\right)$$

3. Topological Relative Persistence

For a dominant generator x_i and a subordinate generator x_j (where $j > i$ in the sorted sequence) residing within a localized spatial bound, we find the intervening saddle point (valley) $v_{i,j}$:

$$v_{i,j} = \inf_{x \in [x_i, x_j]} \tilde{f}(x)$$

Instead of absolute topological persistence (which is vulnerable to amplitude variations), we calculate a **Relative Topological Isolation ratio** $R_{topo} \in [0, 1]$:

$$R_{topo}(j) = \max\left(0,\ 1 - \frac{v_{i,j}}{\tilde{f}(x_j)}\right)$$

If $v_{i,j} \to 0$ (deep valley), $R_{topo} \to 1$ (independent feature).
If $v_{i,j} \to \tilde{f}(x_j)$ (shallow saddle on a dominant shoulder), $R_{topo} \to 0$ (merely a structural perturbation).

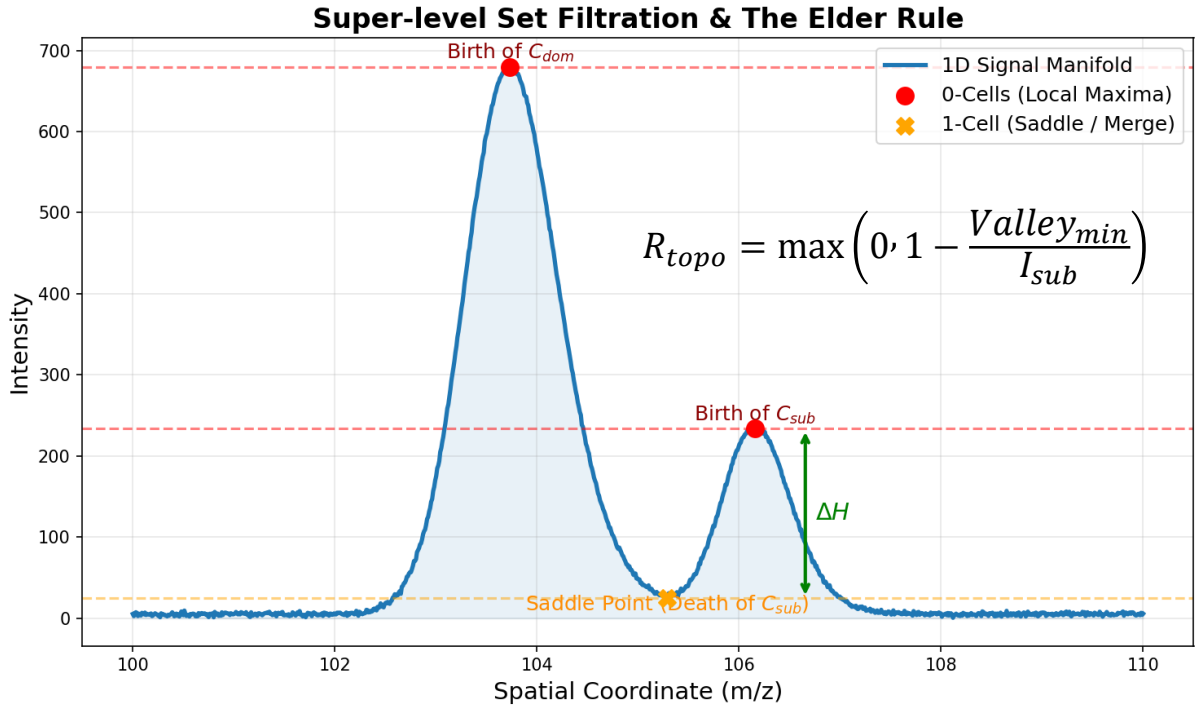
Super-level Set Filtration: Sweeping a threshold from I_{max} to 0 (simulating a "sea-level drop").

0-Dimensional Features (H_0): Local maxima represent the "birth" of new topological components.

The Saddle Point (Index-1 Critical Point): The exact spatial coordinate where two topological components merge.

The Elder Rule Was Applied.

Relative Topological Persistence (R_{topo}): Bounded strictly between $[0, 1]$.



4. Differentiable Persistence Scoring

We map the topological persistence into a probability space via decay parameter α:

$$C_{topo}(j) = 1 - \exp\left(-\alpha \cdot R_{topo}(j)\right)$$

The final confidence metric for any generator j is the product of its physical and topological probabilities. If the subordinate generator is merged into the dominant peak with weak topological isolation, its confidence is penalized:

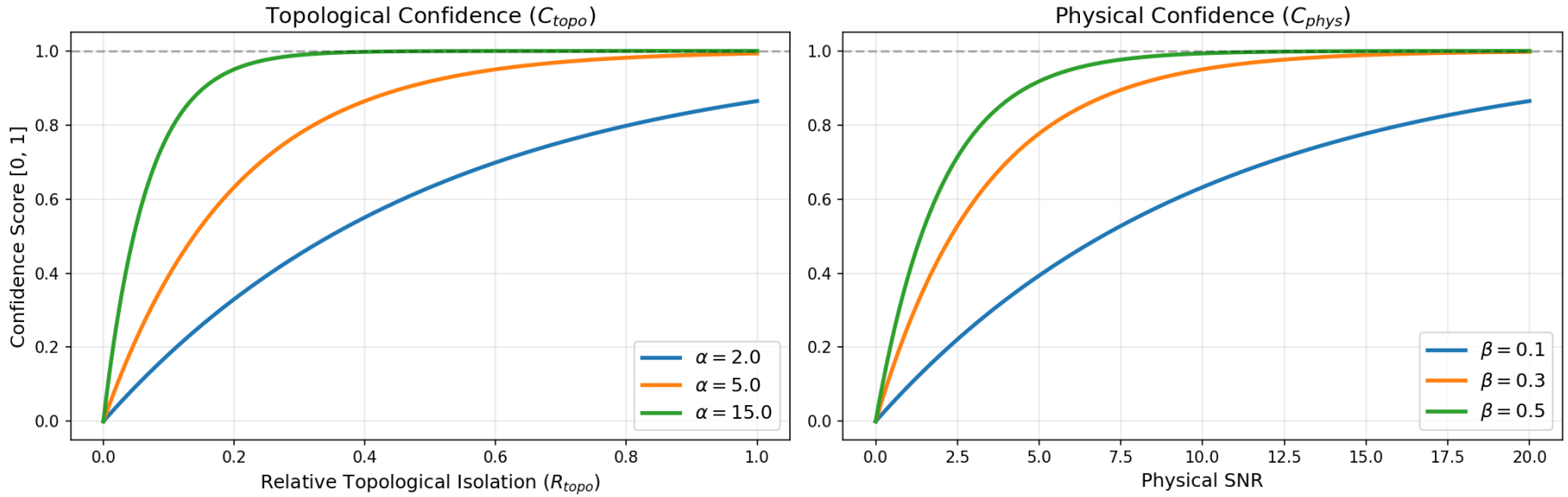
$$C_{final}(j) = \min\left(C_{final}(j),\ C_{topo}(j) \cdot C_{phys}(j)\right)$$

Features where $C_{final} < \tau$ (where τ is the `min_confidence` threshold) are classified as topologically insignificant and discarded.

$$C_{topo} = 1 - \exp(-\alpha \cdot R_{topo})$$

$$C_{phys} = 1 - \exp(-\beta \cdot \mathrm{SNR})$$

$$C_{final} = C_{topo} \times C_{phys}$$

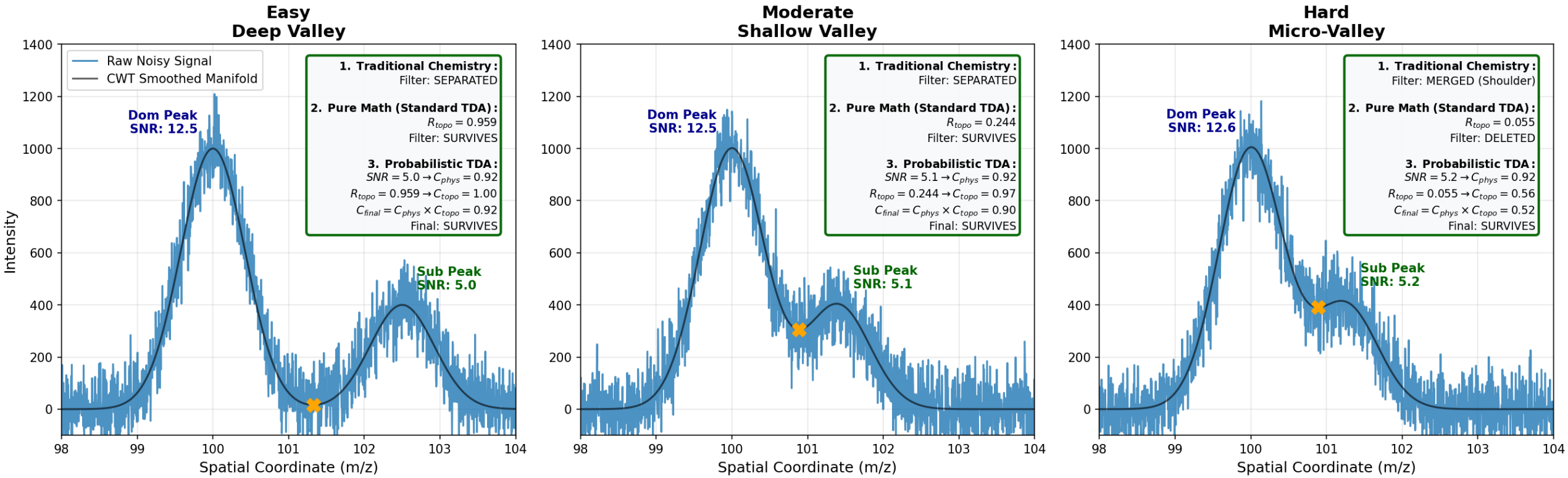


The Noise Dilemma: Standard TDA uses "hard" persistence thresholds, which are brittle against non-linear physical noise (e.g., Poisson shot noise).

Soft Barcodes: We map persistence into a continuous probability space using exponential decay.

Dynamic SNR Calculation: SNR is calculated individually for every single localized peak.

Unifying Math and Physics: High physical SNR can mathematically offset weak topological isolation, and vice versa.



Scale-Space Manifold: Peak topology is evaluated on the CWT-smoothed manifold (black line), effectively peering through raw noise (blue line).

Topological Collapse: In heavy overlap, the Index-1 critical point is submerged by the dominant peak's tail ($R_{topo} < 0.15$). Standard TDA strictly deletes this.

The Rescue Mechanism:

$SNR_{sub} \approx 5 \rightarrow C_{phys} \approx 0.9$. The high physical confidence rescues the heavily decayed topological confidence.

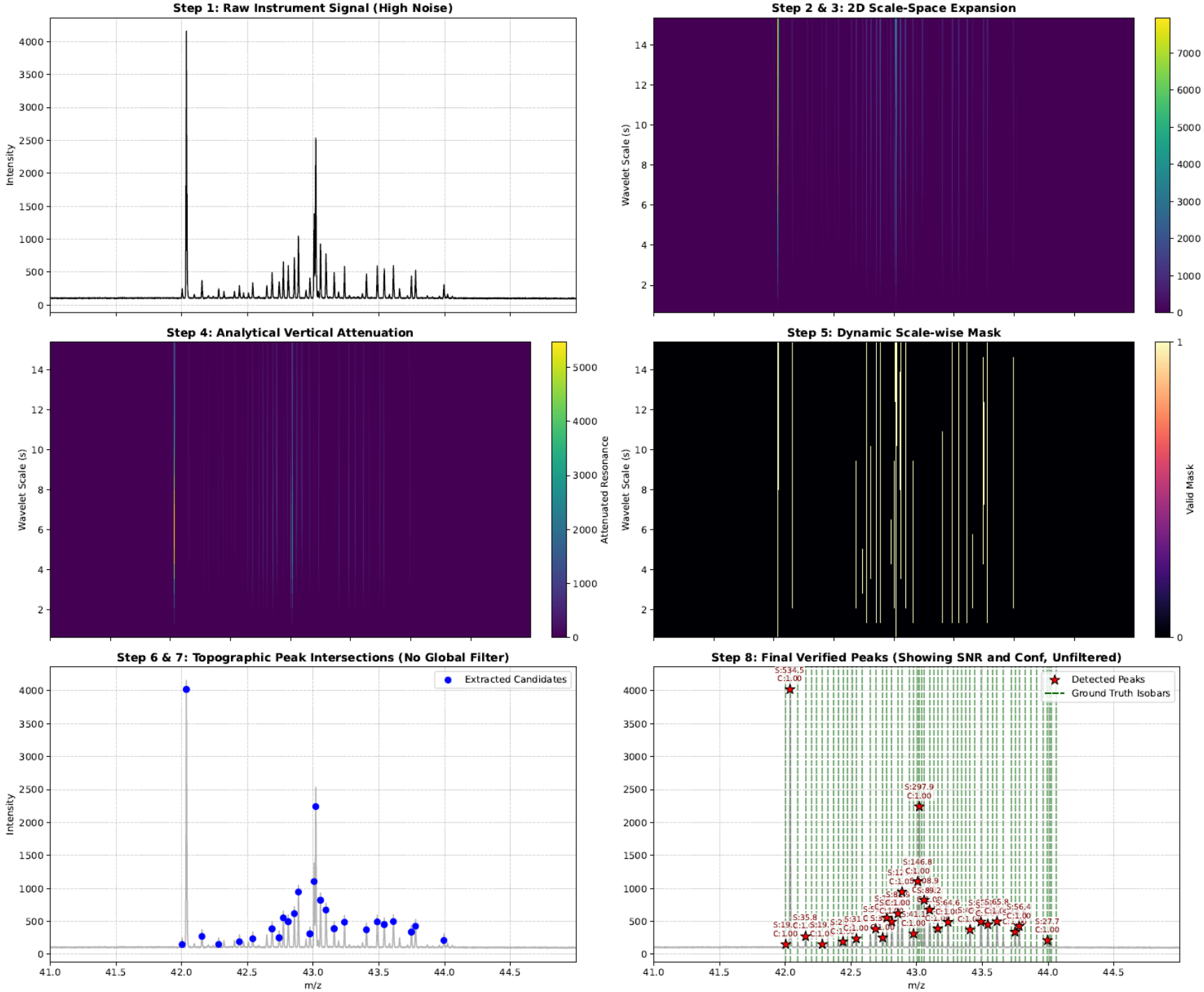
Result: Sub-pixel separation of heavily overlapped components that standard math and chemistry alone would reject.

Phase V: Feature Geometry Integration

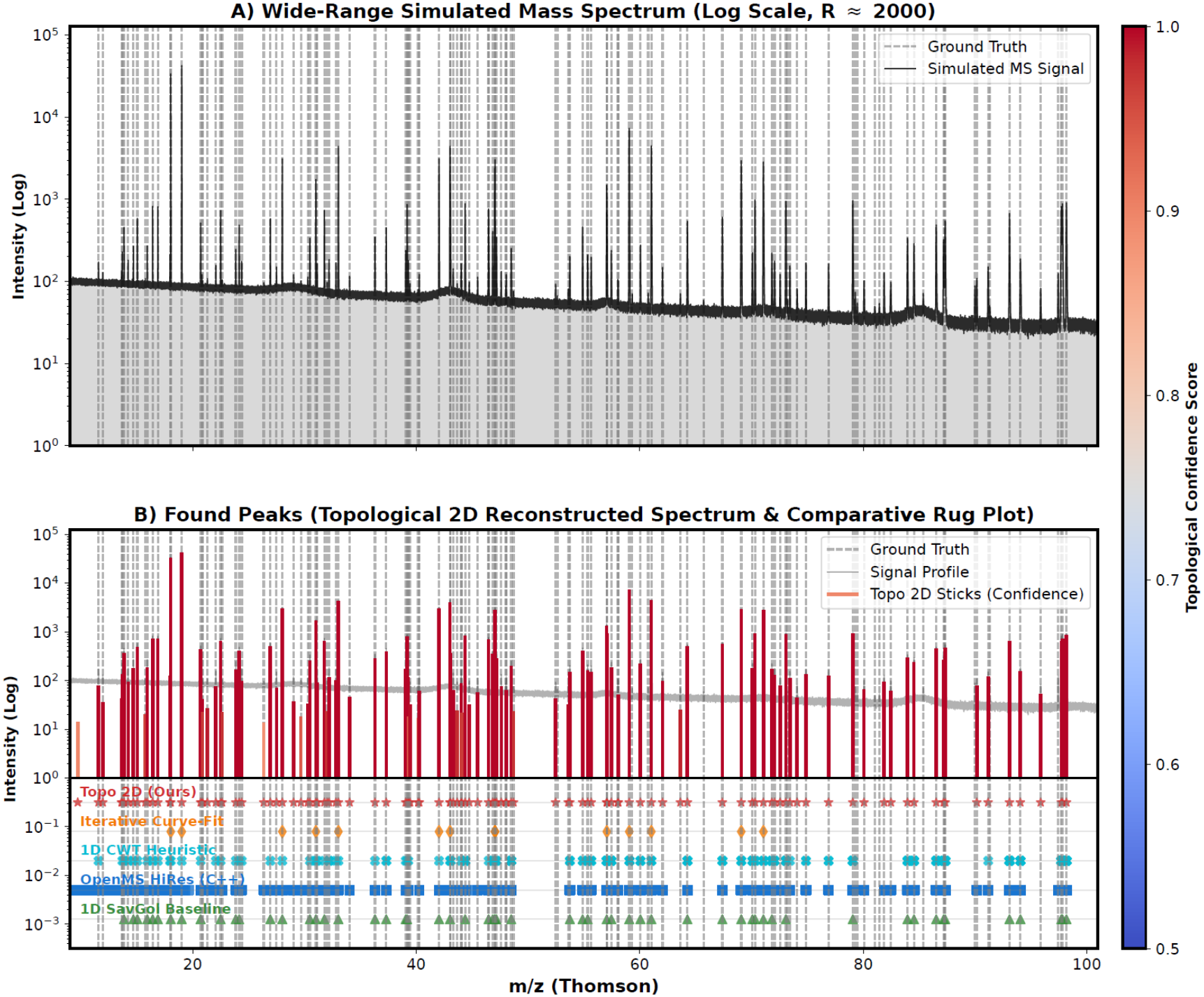
For the set of persisting topological generators $\mathcal{K} \subset \mathcal{M}$, we compute the final integrated geometry (Area).

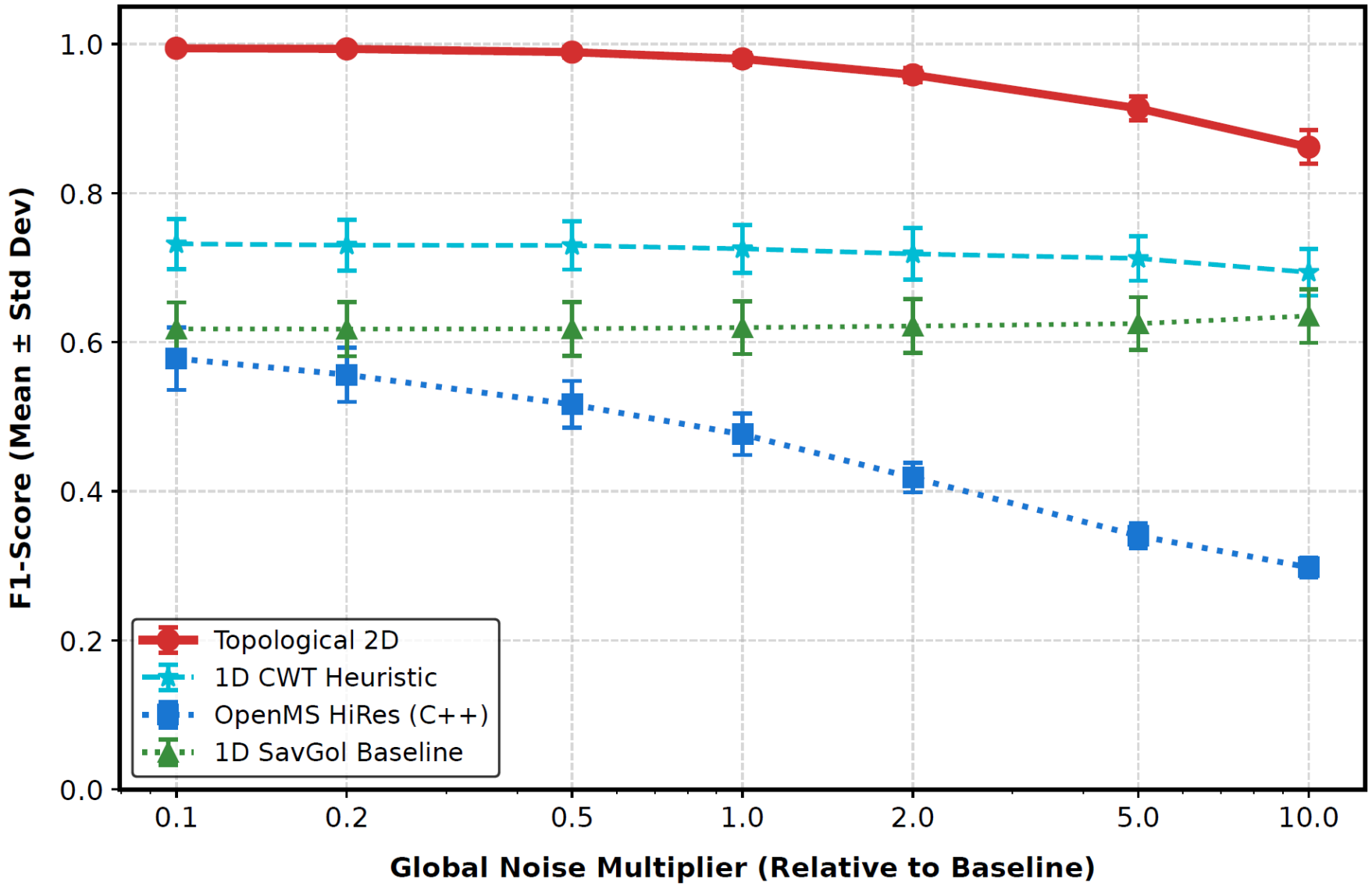
Let $\hat{\sigma}_k = \frac{\hat{s}_k}{2.355}$ be the empirical variance derived from the optimal scale in $\Phi(x, s)$, protected by a conditional fallback to the theoretical variance σ_k at low SNRs.
Applying an asymmetry correction factor γ, the area $\mathcal{A}$ of feature k is estimated via Gaussian integral:

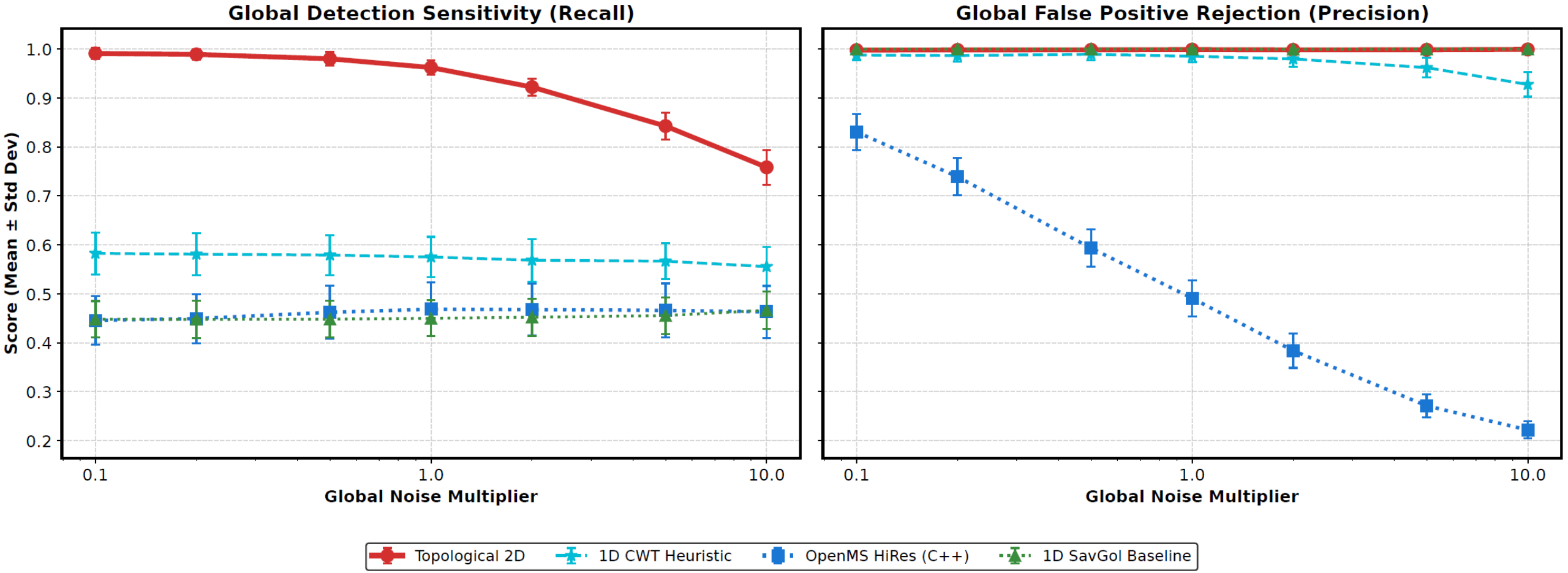
$$\mathcal{A}_k = \tilde{f}(x_k) \cdot \hat{\sigma}_k \sqrt{2\pi} \left(1 + \frac{\gamma}{2}\right)$$

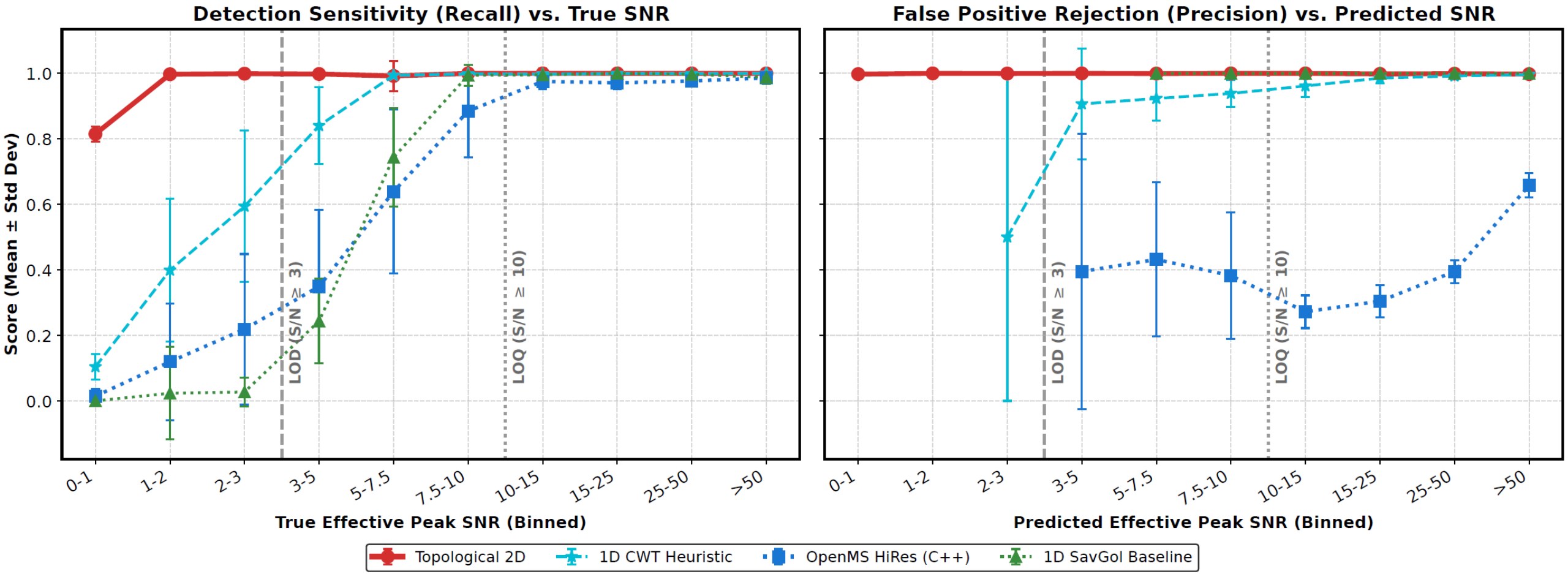

Step 1: Raw Instrument Signal (High Noise)
Intensity
Step 2 & 3: 2D Scale-Space Expansion
Wavelet Scale (s)
Resonance
Step 4: Analytical Vertical Attenuation
Wavelet Scale (s)
Attenuated Resonance
Step 5: Dynamic Scale-wise Mask
Wavelet Scale (s)
Valid Mask
Step 6 & 7: Topographic Peak Intersections (No Global Filter)
Extracted Candidates
Intensity
m/z
Step 8: Final Verified Peaks (Showing SNR and Conf, Unfiltered)
Detected Peaks
Ground Truth Isobars
m/z

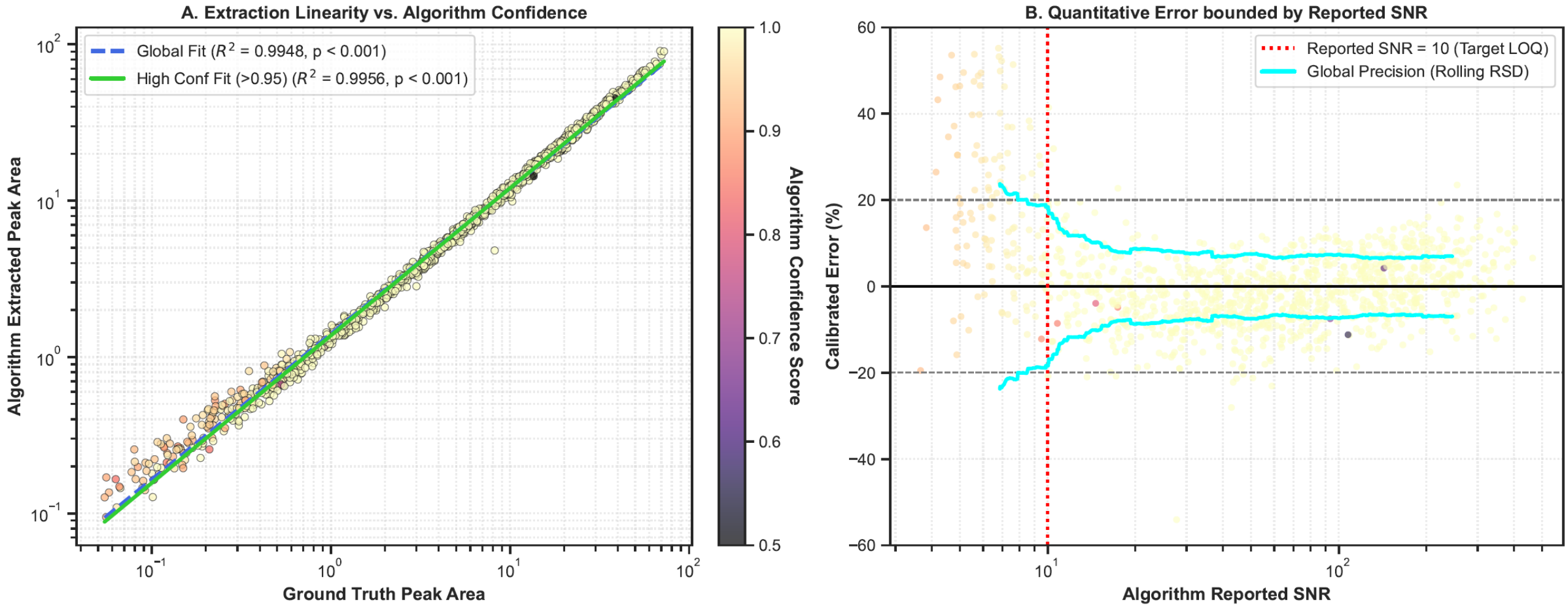
Results


A) Wide-Range Simulated Mass Spectrum (Log Scale, R ≈ 2000)
Intensity (Log)
Ground Truth
Simulated MS Signal
B) Found Peaks (Topological 2D Reconstructed Spectrum & Comparative Rug Plot)
Intensity (Log)
Ground Truth
Signal Profile
Topo 2D Sticks (Confidence)
Topo 2D (Ours)
Iterative Curve-Fit
1D CWT Heuristic
OpenMS HiRes (C++)
1D SavGol Baseline
m/z (Thomson)
Topological Confidence Score

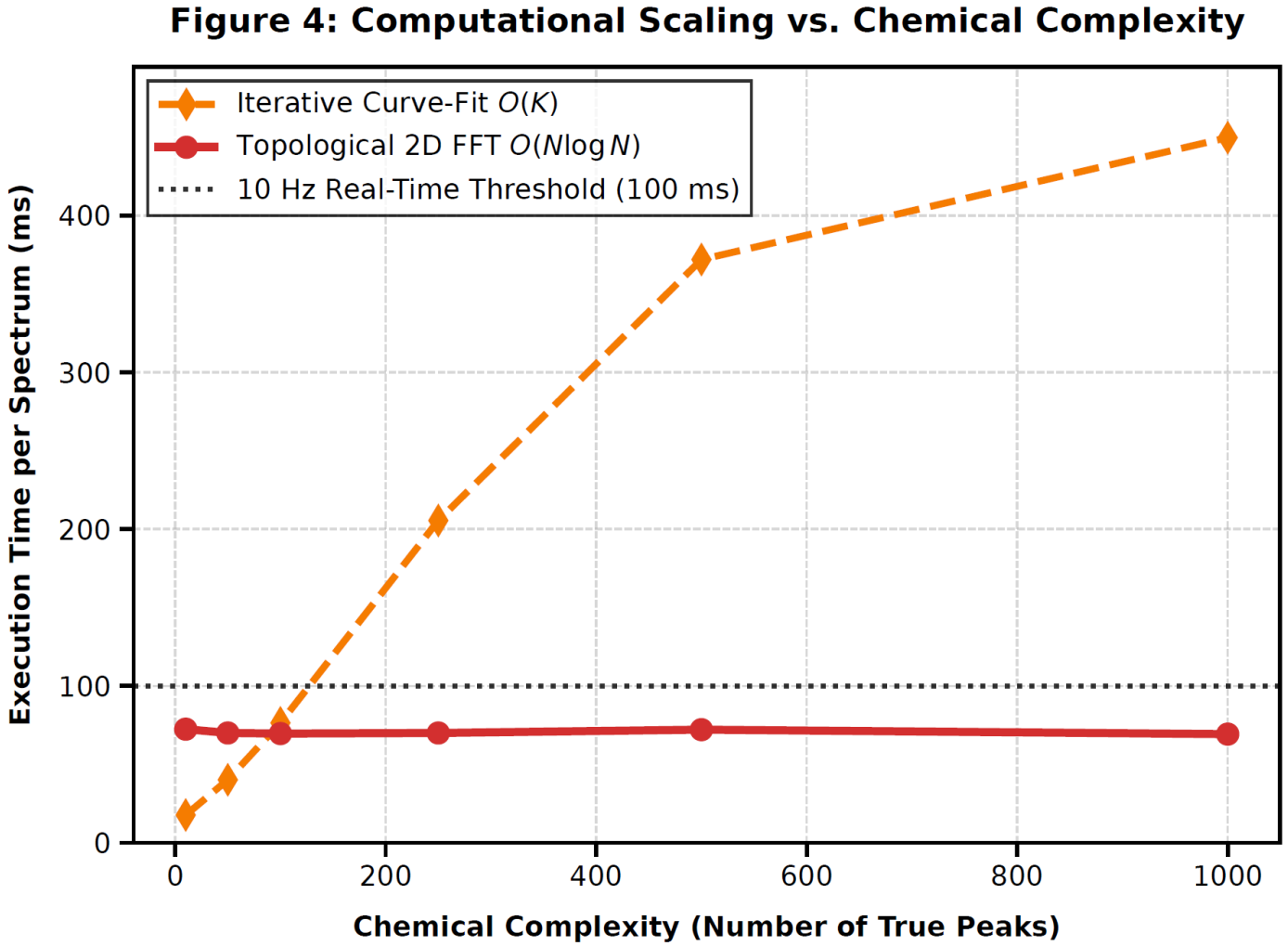

F1-Score (Mean ± Std Dev)
1.0
0.8
0.6
0.4
0.2
0.0
0.1
0.2
0.5
1.0
2.0
5.0
10.0
Global Noise Multiplier (Relative to Baseline)
Topological 2D
1D CWT Heuristic
OpenMS HiRes (C++)
1D SavGol Baseline


Global Detection Sensitivity (Recall)
Global False Positive Rejection (Precision)
Score (Mean ± Std Dev)
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
1.0
10.0
Global Noise Multiplier
Topological 2D
1D CWT Heuristic
OpenMS HiRes (C++)
1D SavGol Baseline


Detection Sensitivity (Recall) vs. True SNR
False Positive Rejection (Precision) vs. Predicted SNR
Score (Mean ± Std Dev)
1.0
0.8
0.6
0.4
0.2
0.0
LOD (S/N ≥ 3)
LOQ (S/N ≥ 10)
0-1
1-2
2-3
3-5
5-7.5
7.5-10
10-15
15-25
25-50
>50
True Effective Peak SNR (Binned)
Predicted Effective Peak SNR (Binned)
Topological 2D
1D CWT Heuristic
OpenMS HiRes (C++)
1D SavGol Baseline


A. Extraction Linearity vs. Algorithm Confidence
Global Fit (R^2 = 0.9948, p < 0.001)
High Conf Fit (>0.95) (R^2 = 0.9956, p < 0.001)
Algorithm Extracted Peak Area
Ground Truth Peak Area
Algorithm Confidence Score
B. Quantitative Error bounded by Reported SNR
Reported SNR = 10 (Target LOQ)
Global Precision (Rolling RSD)
Calibrated Error (%)
Algorithm Reported SNR

Computational Speed & Scalability



•Because our method relies on highly optimized Fast Fourier Transforms (FFT), execution time scales with the array size $O(N \log N)$, not the number of peaks.

•As shown in the graph, iterative curve fitting exponentially slows down as chemical complexity increases. Our algorithm remains practically $O(1)$ relative to complexity, processing 2,500 peaks in the same sub-100ms timeframe as 10 peaks.

Thanks!