Supplementary Information

Identification and Structural Annotation of Unknown Molecular for Miniaturized Mass Spectrometer based the Transformer Enabled Fragment Tree Method

Yiming Yang[1], Shuang Sun[1], Shuyuan Yang[1], Qin Yang[1], Xinqiong Lu[2], Xiaohao Wang[1], Quan Yu[1], Xingming Huo[3,*], Xiang Qian[1,*]

1. Shenzhen International Graduate School, Tsinghua University, Shenzhen, 518055, China
2. Shenzhen Chin Instrument Co. Ltd., Shenzhen, 518055, China
3. School of Biomedical Engineering, Shenzhen Campus of Sun Yat-sen University, Shenzhen 518107, China

* Corresponding Author: qian.xiang@sz.tsinghua.edu.cn, huoxm@mail.sysu.edu.cn

Table S1: Dissociation and rearrangement rules used in the Recap method

Flavanols: "[#6:1]@[#6:2](=[O:3])@[#6:4]>>[#6:1]-[#6:4].[#6H0:2](=[O:3])", "[c:1]-[C:2](=[C:3]-[O:4])-[c:5]:[c:6]:[c:7]:[c:8](-[O:9])@[c:10]@[c:11]>>[c:1]-[C:2]=[C]1[C]=[C]-[C](=[O])-[C:10]=[C:11]1.[C]-[O]", "[c:1]@[Cr4:2]@[Cr4:3]@[c:4]>>[C:2]-[C:3].[c:1]-[c:4]", "[Cr4:1]-[c:2]>>[Cr4:1].[c:2]",
Stilbene: "[c:1]-[C:2]=[C:3]-[c:4]>>[C]=[C]-[C:2]=[C:3]-[c:4]", "[cr6:1]@[cr6:2](-[O:7])@[cr6:3]@[cr6:4]@[cr6:5]@[cr6:6]>>[C:1]-[C:3]-[C:4]=[C:5]-[C:6]", "[C:1]1=[C:3]-[C:4]-[C:5]=[C:6]1>>[C:1]1-[C:3]=[C:6]1",
Commonly rules: "[#7;+0;D2,D3:1]-!@[S:2](=[O:3])=[O:4]>>[#7:1].[S:2](=[O:3])=[O:4]", "[C:1]=!@[C:2]>>[C+:1].[C+:2]", "[#6H2:1]-!@[#6H1:2]-[O;-1]>>[#6H1:1]=,:[#6+:2]", "[#6H2:1]-!@[#6H2:2]-[O;-1]>>[#6H1:1]=,:[#6H1+:2]", "[#6H2:1]-!@[#6H1:2]-[O]>>[#6H1:1]=,:[#6H1:2]", "[#6H2:1]-!@[#6H2:2]-[O]>>[#6H1:1]=,:[#6H2:2]", "[#6H1:1]-!@[#6H1:2]-[O]>>[#6H0:1]=,:[#6H1:2]", "[#6H1:1]-!@[#6H1:2]-[O]>>[#6H0:1]=,:[#6H1:2]", "[C;+0:1]-!@[O;+0,+0:2]>>[C;+0:1].[O;+0,+0:2]", "[c:1]-[c:2](=O)-[c:3]>>[c:1]-[c:3]", "[CH;+0:1]-[OH:2]>>[CH0;+1:1].[O;-1:2]", "[#6;+0:1]-!@[N;+0,+0:2]>>[#6+1:1].[N;+1:2]", "[C;0:1]-!@[O;-1,+0:2]>>[C;+1:1].[O;-1:2]", "[C;0:1]-!@[O;-1,+0:2]>>[C;+1:1].[O;0:2]", "[C;+1:1]-!@[N;+0,+0:2]>>[C;+2:1].[N;+1:2]", "[C;+1:1]-!@[C;+1,+0:2]>>[C;+1:1].[C;+1:2]", "[C;+0:1]-!@[N;+1,+0:2]>>[C;+1:1].[N;+2:2]", "[C;+0:1]-!@[C;+0:2]>>[C;+1:1].[C;+1:2]", "[C;+0:1]-!@[C;+0:2]>>[C;+0:1].[C;+0:2]", "[C;+0:1]-!@[C;+1:2]>>[C;+1:1].[C;+2:2]", "[C;+0:1]-!@[S;+0:2]>>[C;+1:1].[S;+1:2]", "[c;+0:1]-[O;+0:2]>>[c;+0:1].[O;-1:2]", "[c;+0:1]-[C;+0:2]>>[c;+0:1].[C;+0:2]",

S2：Algorithm for Enhancing Mass Spectrometer Resolution

Due to the instability of the equipment hardware and the variability of the operating environment, the system error of the on-site mass spectrometer has significant instability and nonlinearity. This experiment eliminates random errors by averaging multiple spectra to calibrate the left and right drift of the mass spectrum and performs overall calibration of the peak shape through deconvolution.

In Equation 1, the mass spectrometry measurement, denoted as *y*, can be described as the result of a convolution operation between the theoretical isotope distribution, y_0, and the actual peak function, *p*.

$$y = y_0 \otimes p \tag{1}$$

The peak function, *p*, can be mathematically transformed into a well-defined function, denoted as *d*, for instance, through convolution with the filter function, *f*. This filter function serves the purpose of calibrating m/z displacement and rectifying kurtosis distortion, as delineated in Equation 2.

$$d = p \otimes f \tag{2}$$

By organizing these two equations, we can derive Equation 3:

$$y_0 \otimes d = y \otimes f \tag{3}$$

In accordance with Equation 3, given the known theoretical isotope distribution, theoretical peak shape, and the actual mass spectrometry signal, the calibration filter, denoted as f, can be computed through deconvolution. In the realm of mass spectrometers, Gaussian or normal error distributions have been established as reliable approximations for describing the mass function. Consequently, we opt for the Gaussian function as the theoretical peak shape function. This filter is subsequently employed for the calibration of the initial mass spectrometry signal, y, resulting in the calibration mass spectrometry signal, r, as depicted in Equation 4.

$$y \otimes f = r \tag{4}$$

The procedure for computing a calibration filter involves several steps. Initially, a spectrum of a well-defined substance is acquired within the mass spectrometry. Subsequently, according to its chemical formula, the theoretical isotope distribution of this substance is calculated. Finally, deconvolution operations are carried out by employing the Richardson-Lucy algorithm on the actual spectrum of the substance and the theoretical isotope distribution to derive the calibration filter. We employ the Richardson-Lucy algorithm to compute the calibration filter, making an assumption that the noise conforms to a Poisson distribution through a maximum likelihood approach. The iterative principle involves presuming that the sample's probability density function adheres to a Gaussian distribution. Leveraging Bayesian probability, Equation 5 expresses the probability of theoretical mass spectrum occurring while estimating the filter as f.

$$P\big((y_0 * d)|f)\big) = \prod (\frac{(f * y)^{y_0 * d} e^{-f*y}}{y_0 * d}) \tag{5}$$

The maximum value of $P\big((y_0 * d)|f)\big)$ in Equation 5 can be determined by minimizing $-lnP\big((y_0 * d)|f)\big)$, as shown in Equation 6.

$$-lnP\big((y_0 * d)|f)\big) = \sum (-(y_0 * d)\ln(f * y) + f * y) \tag{6}$$

Since $-lnP\big((y_0 * d)|f)\big)$ is a convex function, the minimum value is equivalent to the corresponding value when its derivative is 0, as shown in equation 7, where $y(-i)$ is the flip of $y(i)$.

$$\frac{y_0 * d}{(f^k * y)} * y(-i) = 1 \tag{7}$$

Based on this, the iterative equation for f can be derived, as presented in Equation 8. As the number of iterations k increases, noise may become amplified. In our experimental work, we have selected an iteration number of 20.

$$f^{k+1} = \left\{\frac{y_0 * d}{(f^k * y)} * y(-i)\right\} f^k \tag{8}$$

The calibration filter is applied through convolution to the actual mass spectrum of the substance being measured. This process yields a calibrated spectrum with enhanced resolution.

When calculating the theoretical isotope distribution of substances, according to Rockwood's convolution algorithm based on isotope ratios, the isotope distribution can be calculated by gradually adding the isotope ratio information of a specific atom to the entire molecule. Alternatively, we can represent the isotopic function of atoms as δ functions, associating them with convolution, which entails gradually convolving the isotope function of a specific atom into the entire molecule. According to the time-domain convolution theorem based on Fourier transform, the transformation from time-domain convolution to frequency-domain convolution is equivalent to multiplication. Therefore, the isotope distribution in the mass domain can be converted to the μ domain for processing and subsequently transformed back to the mass domain.

Construct the μ domain function for each element, predicated on its atomic mass and corresponding abundance. As an illustration, consider the element Carbon (C), which exhibits two isotopes, one with a mass of 12 and an abundance of 98.9%, and the other with a mass of 13 and an abundance of 1.1%. The μ domain function for C is delineated in equation 9. By extending this methodology, proceed to deduce the μ domain function, denoted as $f_i(\mu)$, for the diverse elements comprising the molecule.

$$f_c(\mu) = 0.989e^{2\pi i*12\mu} + 0.011e^{2\pi i*13\mu} \tag{9}$$

The μ domain function $f(\mu)$ for the molecular formula can be derived by considering the μ domain functions of each constituent element and the quantity of each element present. Taking CO_2 as an example, there is one carbon atom and two oxygen atoms in this molecule. Based on this information, the μ domain function for the molecular formula CO_2 can be computed, as demonstrated in Equation 10.

$$f_{co_2}(\mu) = f_c(\mu)f_o^2(\mu) \tag{10}$$

Conducting an Inverse Fast Fourier Transform (IFFT) on the peak shape function $S(n)$ in the quality domain yields the peak shape function $s(\mu)$ in the μ domain. Throughout this process, we can extract all necessary data by executing a unilateral Fourier transform on the Gaussian function, leveraging its symmetry and real-number properties. Based on the characteristics of Fourier transforms, the convolution in the mass domain is analogous to multiplication in the μ domain. Consequently, the ultimate μ domain function is presented in Equation 11.

$$\varphi(\mu) = s(\mu)f(\mu) \tag{11}$$

Finally, with respect to $\varphi(\mu)$, after sampling, perform a Fast Fourier Transform (FFT) and convert it to the quality domain, $\psi(m)$. This process allows for the determination of the theoretical isotope distribution of the molecular formula in the mass domain.