nature portfolio

https://doi.org/10.1038/s41592-021-01331-z

Supplementary information

Spectral entropy outperforms MS/MS dot product similarity for small-molecule compound identification

Supplementary Note for

Spectral entropy outperforms MS/MS dot product similarity for small molecule compound identification

Yuanyue Li[(1)], Tobias Kind[(1)], Jacob Folz[(1)], Arpana Vaniya[(1)], Sajjan Singh Mehta[(1,2)] and Oliver Fiehn[(1,*)]

(1) West Coast Metabolomics Center, UC Davis Genome Center, University of California, Davis, 451 Health Sciences Drive, Davis, California 95616, United States
(2) oloBion, Parc Científic de Barcelona, Avenida Dr. Marañón 8, 08028 Barcelona, Spain
(*) Corresponding authors Emails: ofiehn@ucdavis.edu

Supplementary Note 1: “Equations for Calculating MS/MS Similarity”

We define m/z as M, and intensity as I.

The test spectrum q is a collection of peaks: $\left(M_{q,1}, I_{q,1}\right), \ldots, \left(M_{q,i}, I_{q,i}\right)$, the library spectrum r is defined as $\left(M_{r,1}, I_{r,1}\right), \ldots, \left(M_{r,i}, I_{r,i}\right)$.

$N_q, N_r$is the number of fragment ions for spectrum q, r; N_m is the total number of matching fragment ions.

Thus, we have the average intensity $\overline{I_q} = \frac{\sum I_{q,i}}{N_q}, \overline{I_r} = \frac{\sum I_{r,i}}{N_r}$.

Here ln is the natural logarithm log_e.

The spectral similarity is calculated with follow equations:

Unweighted entropy similarity: $-\frac{2\times S_{AB}-S_A-S_B}{\ln(4)}, S = -\sum_p I_p \ln I_p$

Entropy similarity: $1-\frac{2\times S'_{AB}-S'_A-S'_B}{\ln(4)}, S' = -\sum_p I'_p \ln I'_p\,, with \begin{cases} I' = I \ (S \geq 3) \\ I' = I^w, w = 0.25 + S * 0.25 \ (S < 3) \end{cases}$

Dot-product (cosine) similarity: $\frac{\left(\sum I_{q,i} \cdot I_{r,i}\right)^2}{\sum I_{q,i}^2 \cdot \sum I_{r,i}^2}$

Weighted dot product similarity: $\frac{\left(\sum W_{q,i} \cdot W_{r,i}\right)^2}{\sum W_{q,i}^2 \cdot \sum W_{r,i}^2}, W_i = M_i^3 I_i^{0.6}$

Reverse dot-product similarity: $\frac{\left(\sum I'_{q,i} \cdot I'_{r,i}\right)^2}{\sum I'^2_{q,i} \cdot \sum I'^2_{r,i}} with\ I'_{q,i} > 0, I'_{r,i} > 0\ and\ M'_{q,i} = M'_{r,i}$

MSforID similarity version 1: $-\frac{N_m^4}{N_q N_r \left(\sum\left|I_{q,i}-I_{r,i}\right|\right)^{0.25}}$

MSforID similarity: $-\frac{N_m^4 \left(\sum I_{q,i}+2\sum I_{r,i}\right)^{1.25}}{\left(N_q+2N_r\right)^2+\sum\left|I_{q,i}-I_{r,i}\right|+\sum\left|M_{q,i}-M_{r,i}\right|}$

Euclidean similarity: $1-\frac{1}{\sqrt{2}}\sqrt{\sum_i\left(I_{q,i}-I_{r,i}\right)^2}$

Manhattan similarity: $1-\frac{1}{2}\sum_i\left|I_{q,i}-I_{r,i}\right|$

Chebyshev similarity: $1-\max_i\left(\left|I_{q,i} - I_{r,i}\right|\right)$

Squared euclidean similarity: $1-\frac{1}{2}\sum_i\left(I_{q,i}-I_{r,i}\right)^2$

Fidelity similarity: $\sum\sqrt{I_{q,i} \cdot I_{r,i}}$

Matusita similarity: $1-\frac{1}{2}\sqrt{\sum\left(\sqrt{I_{q,i}}-\sqrt{I_{r,i}}\right)^2}$

Squared-chord similarity: $1-\frac{1}{2}\sum\left(\sqrt{I_{q,i}}-\sqrt{I_{r,i}}\right)^2$

Bhattacharya 1 similarity: $1-\frac{1}{(\cos^{-1}0)^2}\left(\cos^{-1}\left(\sum\sqrt{I_{q,i}\cdot I_{r,i}}\right)\right)^2$

Bhattacharya 2 similarity: $\frac{1}{1-\ln\left(\sum\sqrt{I_{q,i}\cdot I_{r,i}}\right)}$

Harmonic mean similarity: $2\cdot\sum\left(\frac{I_{q,i}I_{r,i}}{I_{q,i}+I_{r,i}}\right)$

Probabilistic symmetric χ^2 similarity: $1-2\cdot\sum\frac{\left(I_{q,i}-I_{r,i}\right)^2}{I_{q,i}+I_{r,i}}$

Ruzicka similarity: $1-\frac{\sum|I_{q,i}-I_{r,i}|}{\sum\max\,(I_{q,i},I_{r,i})}$

Roberts similarity: $\sum\frac{(I_{q,i}+I_{r,i})}{\sum(I_{q,i}+I_{r,i})}\frac{\min(I_{q,i},I_{r,i})}{\max(I_{q,i},I_{r,i})}$

Intersection similarity: $\frac{\sum\min(I_{q,i},I_{r,i})}{\min\,(\sum I_{q,i},\sum I_{r,i})}$

Motyka similarity: $2\cdot\frac{\sum\min(I_{q,i},I_{r,i})}{\sum(I_{q,i}+I_{r,i})}$

Canberra similarity: $1-\frac{1}{1+\sum\frac{\left|I_{q,i}-I_{r,i}\right|}{\left|I_{q,i}\right|+\left|I_{r,i}\right|}}$

Baroni-Urbani-Buser similarity: $\frac{\sum I_{i,min}+\sqrt{\sum(I_{i,min}\cdot\sum(I_{max}-I_{i,max}))}}{\sum I_{i,max}+\sqrt{\sum(I_{i,min}\cdot\sum(I_{max}-I_{i,max}))}}, with \begin{cases} I_{i,min}=\min(I_{q,i},I_{r,i}) \\ I_{i,max}=\max(I_{q,i},I_{r,i}) \\ I_{max}=\max(I_q,I_r) \end{cases}$

Penrose size similarity: $\frac{1}{1+\sqrt{N_q}\sum|I_{q,i}-I_{r,i}|}$

Mean character similarity: $1-\frac{1}{2N_q}\sum|I_{q,i}-I_{r,i}|$

Lorentzian similarity: $\frac{1}{1+\sum\ln\,(1+|I_{q,i}-I_{r,i}|)}$

Penrose shape similarity: $1-\frac{1}{\sqrt{2}}\sqrt{\sum\left(\left(I_{q,i}-\overline{I_q}\right)-\left(I_{r,i}-\overline{I_r}\right)\right)^2}$

Clark similarity: $\frac{1}{1+\sqrt{\frac{1}{N_q}\sum\left(\frac{I_{q,i}-I_{r,i}}{|I_{q,i}|+|I_{r,i}|}\right)^2}}$

Hellinger similarity: $\frac{1}{1+\sqrt{2\sum\left(\sqrt{\frac{I_{q,i}}{\overline{I_q}}}-\sqrt{\frac{I_{r,i}}{\overline{I_r}}}\right)^2}}$

Whittaker index of association similarity: $\frac{1}{1+\frac{1}{2}\sum\left|\frac{I_{q,i}}{\overline{I_q}}-\frac{I_{r,i}}{\overline{I_r}}\right|}$

Symmetric χ2 similarity: $1-\sqrt{2\cdot\sum\frac{\overline{I_q}+\overline{I_r}}{N_q\left(\overline{I_q}+\overline{I_r}\right)^2}\frac{\left(I_{q,i}\cdot\overline{I_r}-I_{r,i}\cdot\overline{I_q}\right)^2}{I_{q,i}+I_{r,i}}}$

Pearson/Spearman similarity: $\frac{1}{2}\left(1+\frac{\sum\left[\left(I_{q,i}-\overline{I_q}\right)\left(I_{r,i}-\overline{I_r}\right)\right]}{\sqrt{\sum\left(I_{q,i}-\overline{I_q}\right)^2\sum\left(I_{r,i}-\overline{I_r}\right)^2}}\right)$

Improved similarity: $\frac{1}{1+\sqrt{\frac{1}{N_q}\sum\left(\frac{I_{q,i}-I_{r,i}}{I_{q,i}+I_{r,i}}\right)^2}}$

Absolute Value similarity: $\frac{1}{1+\frac{\sum\left(\left|I_{q,i}-I_{r,i}\right|\right)}{\sum I_{q,i}}}$

Spectral contrast angle similarity: $\frac{\sum I_{q,i}\cdot I_{r,i}}{\sqrt{\sum I_{q,i}^2\cdot\sum I_{r,i}^2}}$

Wave Hedges similarity: $1-\sum\frac{\left|I_{q,i}-I_{r,i}\right|}{\max\left(I_{q,i},I_{r,i}\right)}$

Jaccard similarity: $1-\frac{\sum\left(I_{q,i}-I_{r,i}\right)^2}{\sum I_{q,i}^2+\sum I_{r,i}^2-\sum I_{q,i}\cdot I_{r,i}}$

Dice similarity: $1-\frac{\sum\left(I_{q,i}-I_{r,i}\right)^2}{\sum I_{q,i}^2+\sum I_{r,i}^2}$

Inner product similarity: $\sum I_{q,i}\cdot I_{r,i}$

Divergence similarity: $\frac{1}{1+2\sum\frac{\left(I_{q,i}-I_{r,i}\right)^2}{\left(I_{q,i}+I_{r,i}\right)^2}}$

Avg (L1, L∞) similarity: $1-\frac{1}{3}\left(\sum\left|I_{q,i}-I_{r,i}\right|+\max\left|I_{q,i}-I_{r,i}\right|\right)$

Vicis-Symmetric χ2 3 similarity: $1-\frac{1}{2}\sum\frac{\left(I_{q,i}-I_{r,i}\right)^2}{\max\left(I_{q,i},I_{r,i}\right)}$

Supplementary Note 2: "Equations for entropy similarity"

We define m/z as M, and intensity as I. A spectrum A is a collection of peaks: $(M_{A,1}, I_{A,1}), \dots, (M_{A,i}, I_{A,i})$.

For calculating the entropy similarity between a spectrum A and a spectrum B defined as $(M_{B,1}, I_{B,1}), \dots, (M_{B,i}, I_{B,i})$, we generate a combined spectrum AB with:

$$I_{AB,i} = \frac{1}{2} I_{A,i} + \frac{1}{2} I_{B,i}$$

By the definition of spectral entropy, we have:

$$S = -\sum_i I_i \ln I_i$$

Hence, the entropy distance is calculated by:

$$\begin{aligned}
&2 \times S_{AB} - S_A - S_B \\
&= -\left(\sum_i (2 \times I_{AB,i} \ln I_{AB,i}) - \sum_i I_{A,i} \ln I_{A,i} - \sum_i I_{B,i} \ln I_{B,i}\right) \\
&= -\sum_i \left((I_{A,i} + I_{B,i}) \ln I_{AB,i} - I_{A,i} \ln I_{A,i} - I_{B,i} \ln I_{B,i}\right) \\
&= \sum_i \left(I_{A,i} \ln \frac{I_{A,i}}{I_{AB,i}} + I_{B,i} \ln \frac{I_{B,i}}{I_{AB,i}}\right) \\
&= \sum_i \left(I_{A,i} \ln \frac{2 \times I_{A,i}}{I_{A,i} + I_{B,i}} + I_{B,i} \ln \frac{2 \times I_{B,i}}{I_{A,i} + I_{B,i}}\right)
\end{aligned}$$

If spectra A and B are identical, we have $I_{A,i} = I_{B,i}$, and the entropy distance becomes zero:

$$2 \times S_{AB} - S_A - S_B = \sum_i \left(I_{A,i} \ln \frac{2 \times I_{A,i}}{I_{A,i} + I_{A,i}} + I_{A,i} \ln \frac{2 \times I_{A,i}}{I_{A,i} + I_{A,i}}\right) = 0$$

If there are no common fragment ions found between spectra A and B, the entropy distance becomes maximal:

$$\begin{aligned}
&2 \times S_{AB} - S_A - S_B = \sum_i \left(I_{A,i} \ln \frac{2 \times I_{A,i}}{I_{A,i} + I_{B,i}} + I_{B,i} \ln \frac{2 \times I_{B,i}}{I_{A,i} + I_{B,i}}\right) \\
&= \sum_i \left(I_{A,i} \ln \frac{2 \times I_{A,i}}{I_{A,i}}\right) + \sum_j \left(I_{B,j} \ln \frac{2 \times I_{B,j}}{I_{B,j}}\right) \\
&= \ln 2 \times \sum_i I_{A,i} + \ln 2 \times \sum_j I_{B,j} = \ln 2 + \ln 2 = \ln 4
\end{aligned}$$

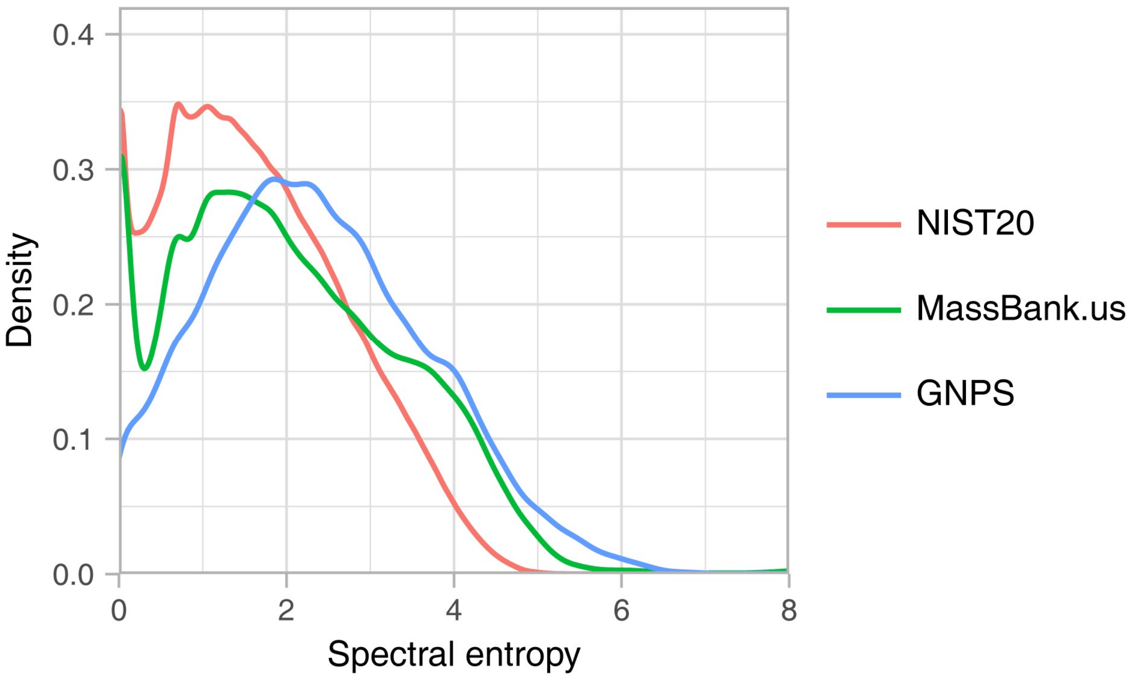
Therefore, the entropy distance ranges from 0 to $\ln 4$.

We then defined the unweighted spectral entropy similarity by normalizing the entropy distance to $[0,1]$.

We obtain the unweighted spectral entropy similarity:

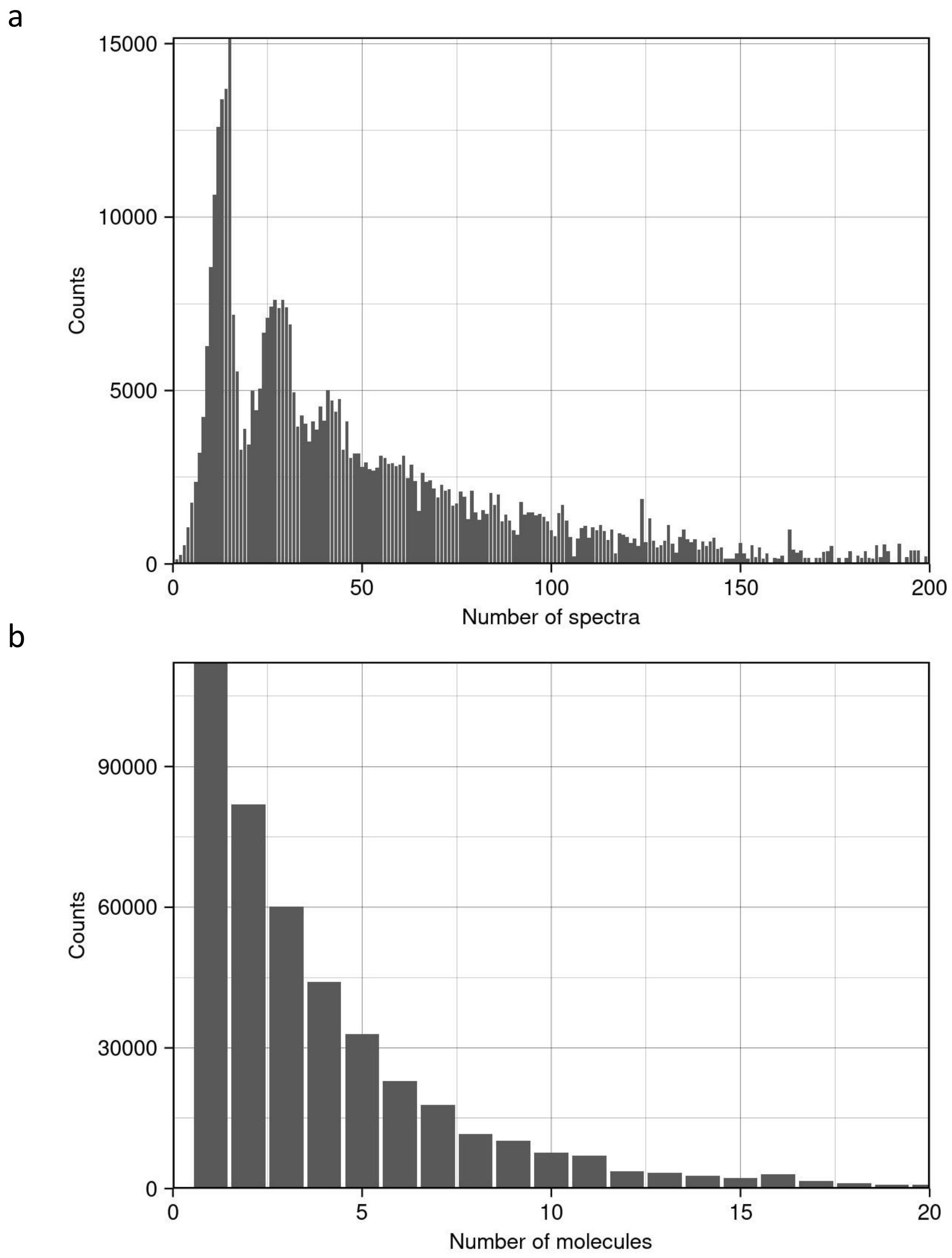
$$1-\frac{2\times S'_{AB}-S'_{A}-S'_{B}}{\ln(4)}$$

Supplementary Figure 1



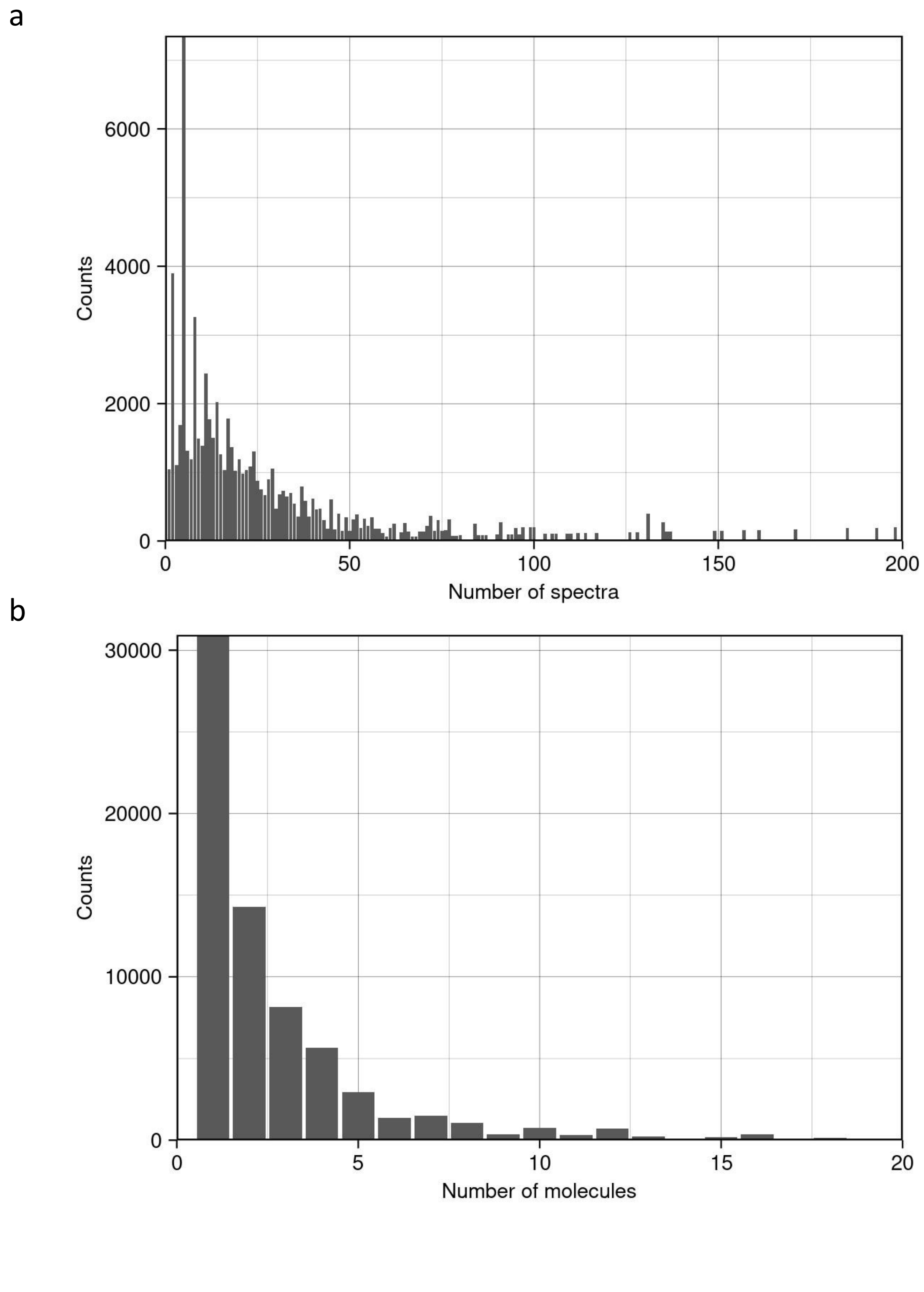
Supplementary Figure 1.
Distribution of spectral entropy values of all spectra in MassBank.us, NIST20, and the GNPS database after removing noise peaks defined as less than 1% base peak intensity.

Supplementary Figure 2



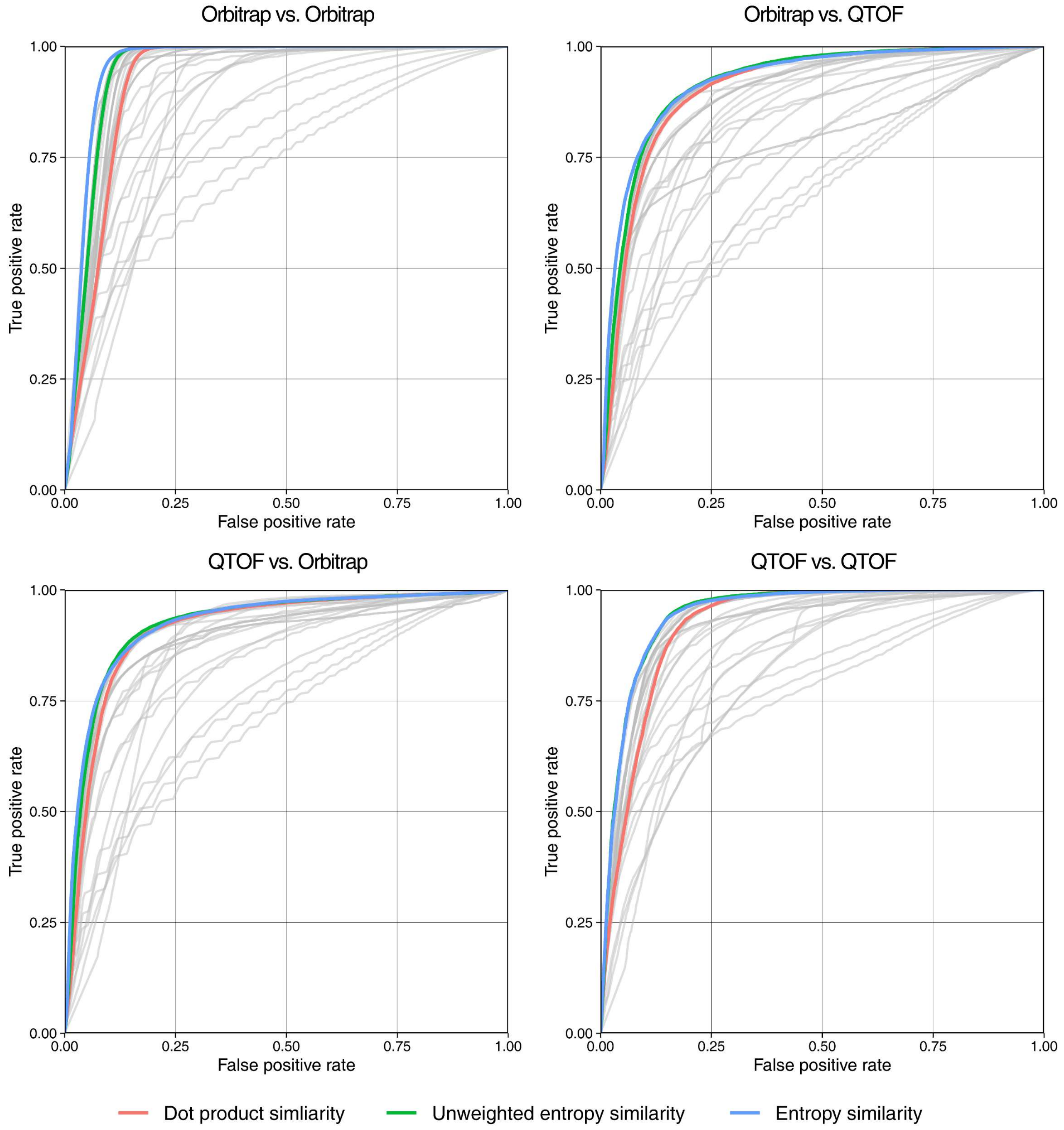
Supplementary Figure 2.
The distribution of candidate hits for NIST20 similarity algorithm benchmarking tests.
(a) Number of candidate spectra per test spectrum.
(b) Number of candidate molecules per test spectrum.

Supplementary Figure 3



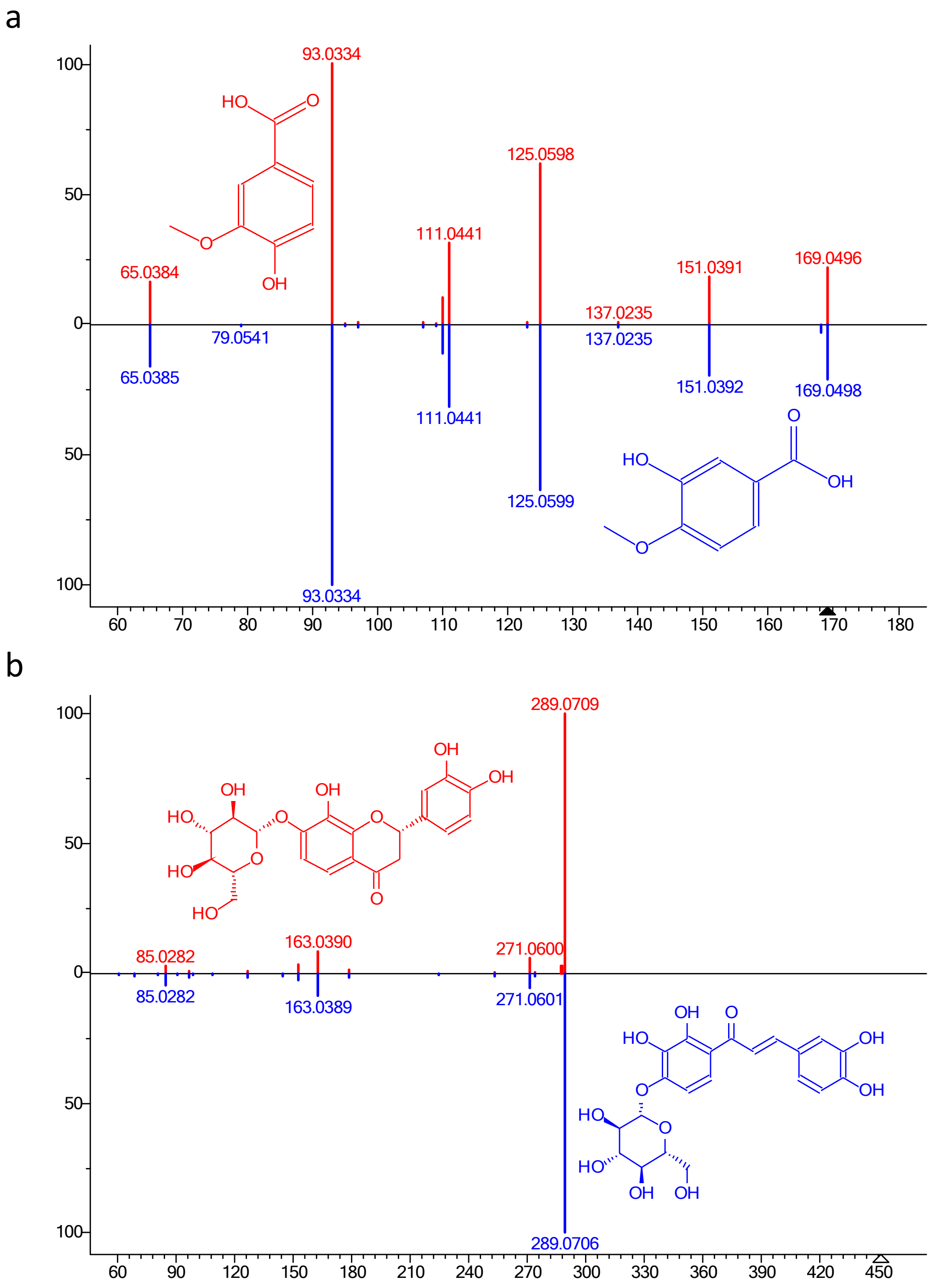
Supplementary Figure 3.
The distribution of candidate hits for Massbank.us similarity algorithm benchmarking tests.
(a) Number of candidate spectra per test spectrum.
(b) Number of candidate molecules per test spectrum.

Supplementary Figure 4



Supplementary Figure 4.
Receiver-operator characteristic curves for all algorithms when searching NIST20 MS/MS spectra separated by type of mass spectrometer. The entropy similarity, unweighted entropy similarity and dot product similarity are highlighted. All other methods are shown in grey.

Supplementary Figure 5



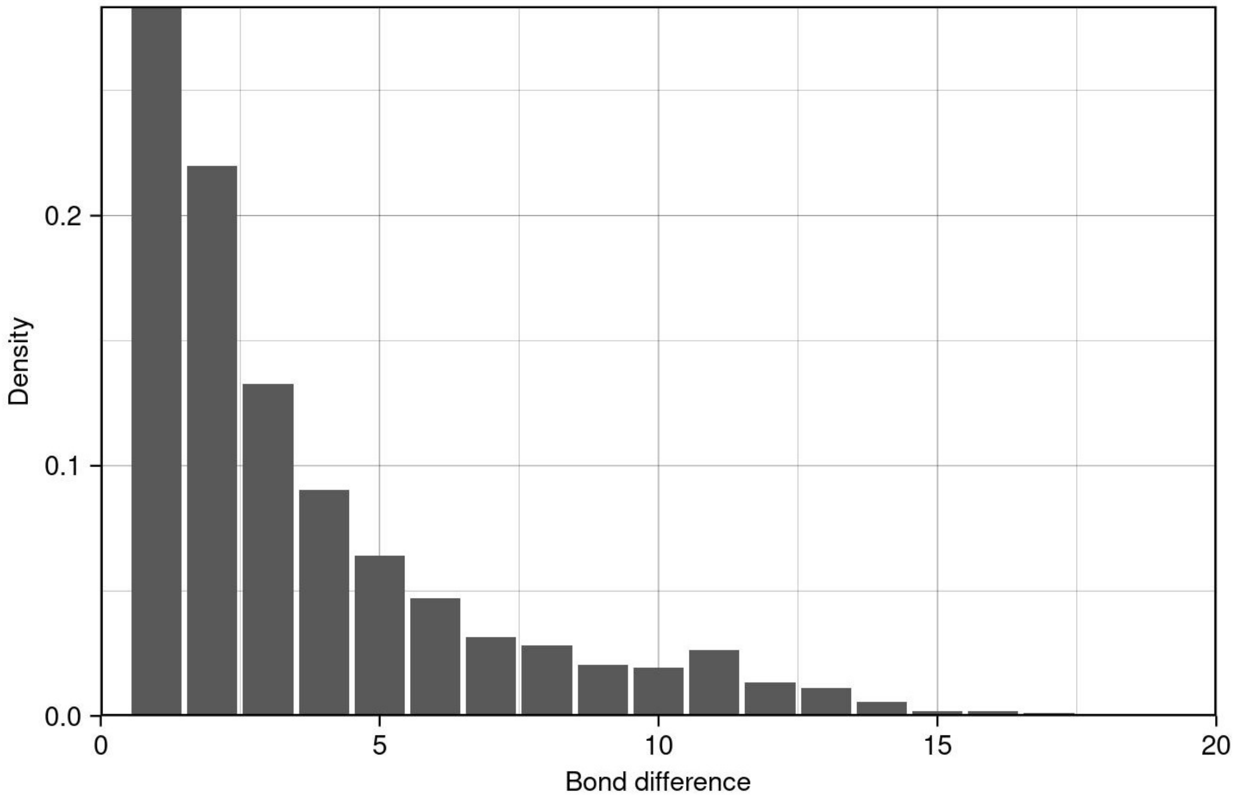
Supplementary Figure 5.

Examples of MS/MS similarities of isomeric molecules.

(a) Comparison of structures and MS/MS spectra for vanillic acid (top) and 3-Hydroxy-4-methoxybenzoic acid (bottom) as example of one-bond different isomers.

(b) Comparison of structures and MS/MS spectra for Flavanomarein (top) and Marein (bottom) as example of multiple-bond different isomers

Supplementary Figure 6



Supplementary Figure 6.
The distribution of bond difference in misidentified molecular pairs.