Flash entropy search to query all mass spectral libraries in real time

Yuanyue Li[1] & Oliver Fiehn[1]

[1]UC Davis West Coast Metabolomics Center, University of California, Davis, CA

UC DAVIS West Coast Metabolomics Center

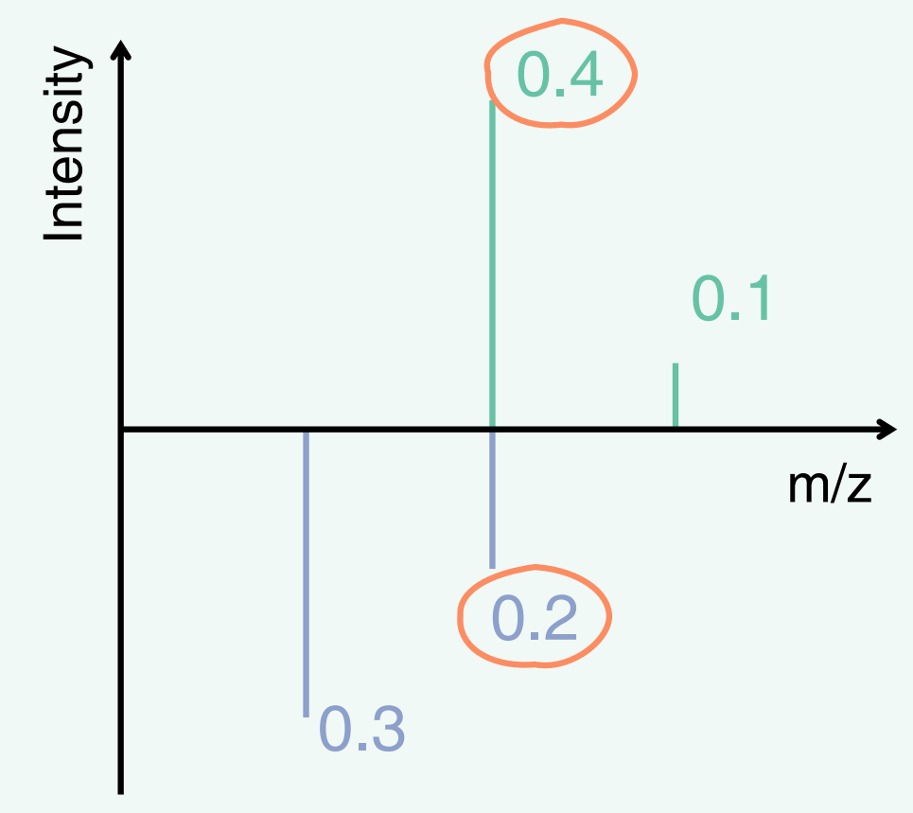
Abstract

Public repositories of metabolomics mass spectra encompass more than 1 billion entries. With open search, dot product or entropy similarity, comparisons of a single tandem mass spectrometry spectrum take more than 8 h. Flash entropy search speeds up calculations more than 10,000 times to query 1 billion spectra in less than 2 s, without loss in accuracy. It benefits from using multiple threads and GPU calculations. This algorithm can fully exploit large spectral libraries with little memory overhead for any mass spectrometry laboratory.

New formula to calculate entropy similarity

$$Entropy\ similarity = \sum f(I_a + I_b) - f(I_a) - f(I_b)$$

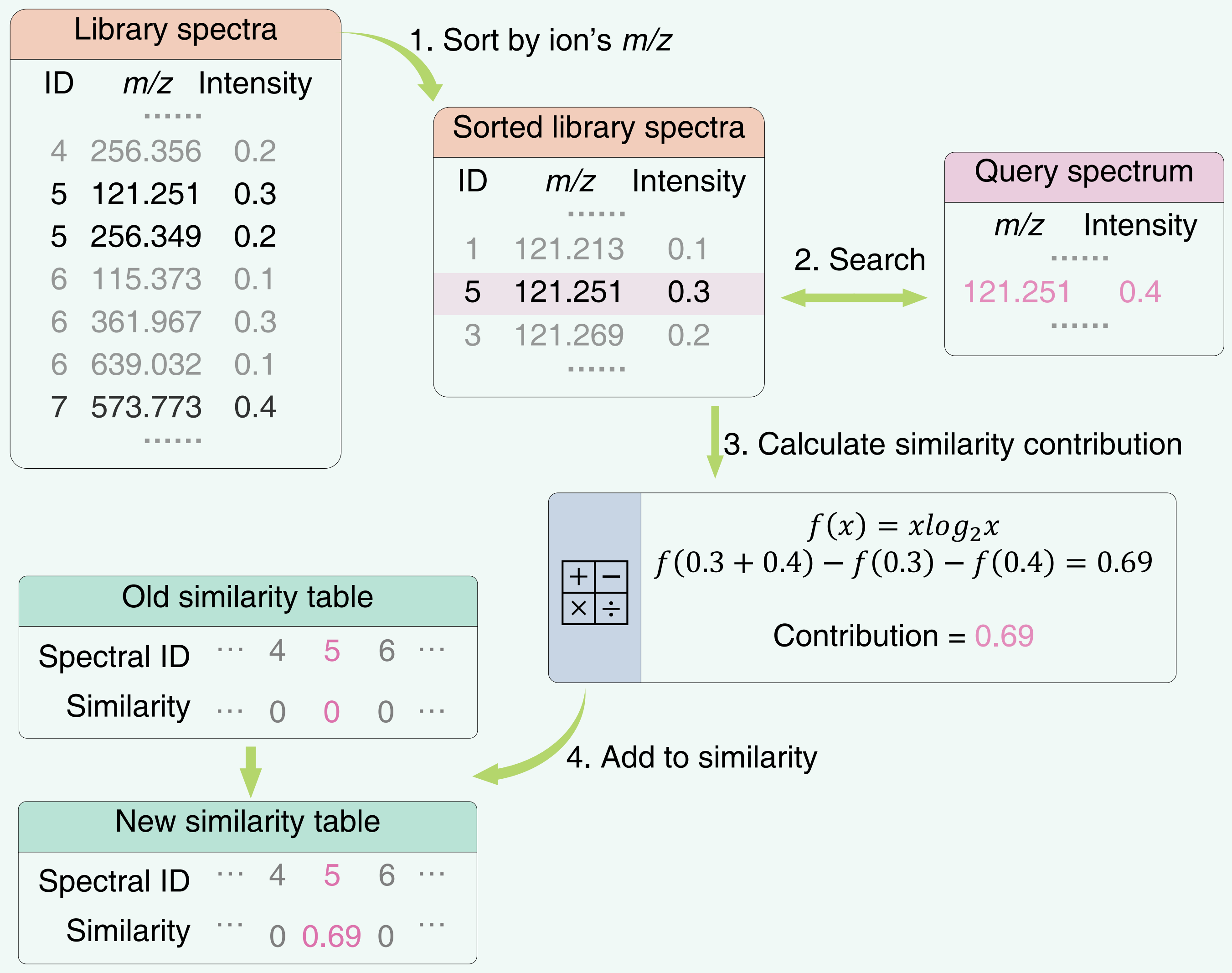
$$f(x) = x \log_2 x$$



$$
\begin{aligned}
&Entropy\ similarity\\
&= f(0.4 + 0.2) - f(0.4) - f(0.2)\\
&= f(0.6) - f(0.4) - f(0.2)\\
&= 0.6 \times \log_2 0.6 - 0.4 \times \log_2 0.4 - 0.2 \times \log_2 0.2\\
&= 0.55
\end{aligned}
$$

Figure 1: Example of calculating entropy similarity. Note that the sum intensities of ion abundances in each spectrum are normalized to equal 0.5

Flash entropy search algorithm



Library spectra

ID	m/z	Intensity
	......	
4	256.356	0.2
5	121.251	0.3
5	256.349	0.2
6	115.373	0.1
6	361.967	0.3
6	639.032	0.1
7	573.773	0.4
	......	

1. Sort by ion's *m/z*

Sorted library spectra

ID	m/z	Intensity
	......	
1	121.213	0.1
5	121.251	0.3
3	121.269	0.2
	......	

2. Search

Query spectrum

m/z	Intensity
......	
121.251	0.4
......	

3. Calculate similarity contribution

$f(x) = xlog_2x$

$f(0.3 + 0.4) - f(0.3) - f(0.4) = 0.69$

Contribution = 0.69

Old similarity table

Spectral ID	···	4	5	6	···
Similarity	···	0	0	0	···

4. Add to similarity

New similarity table

Spectral ID	···	4	5	6	···
Similarity	···	0	0.69	0	···

Figure 2: Workflow for the Flash entropy search algorithm. Spectra are cleaned and normalized to $\sum_i I_i = 0.5$. All library fragment ions are sorted by m/z. Query spectra are used to look up library spectra with matching fragment ions within Δm/z = 20 mDa. Subsequently, entropy similarity contributions are calculated only for these matching ions, greatly enhancing the overall search speed. Finally, this similarity contribution is added to the similarity table for each library query.

Use Flash entropy search

Via programming:
MS Entropy package

Via GUI:
Entropy Search

Fast

- Flash entropy search is more than 10,000 times faster than traditional methods.
- Flash entropy search can perform open search against 1 million spectra in 1 million second, and 937 million spectra in 2 seconds.

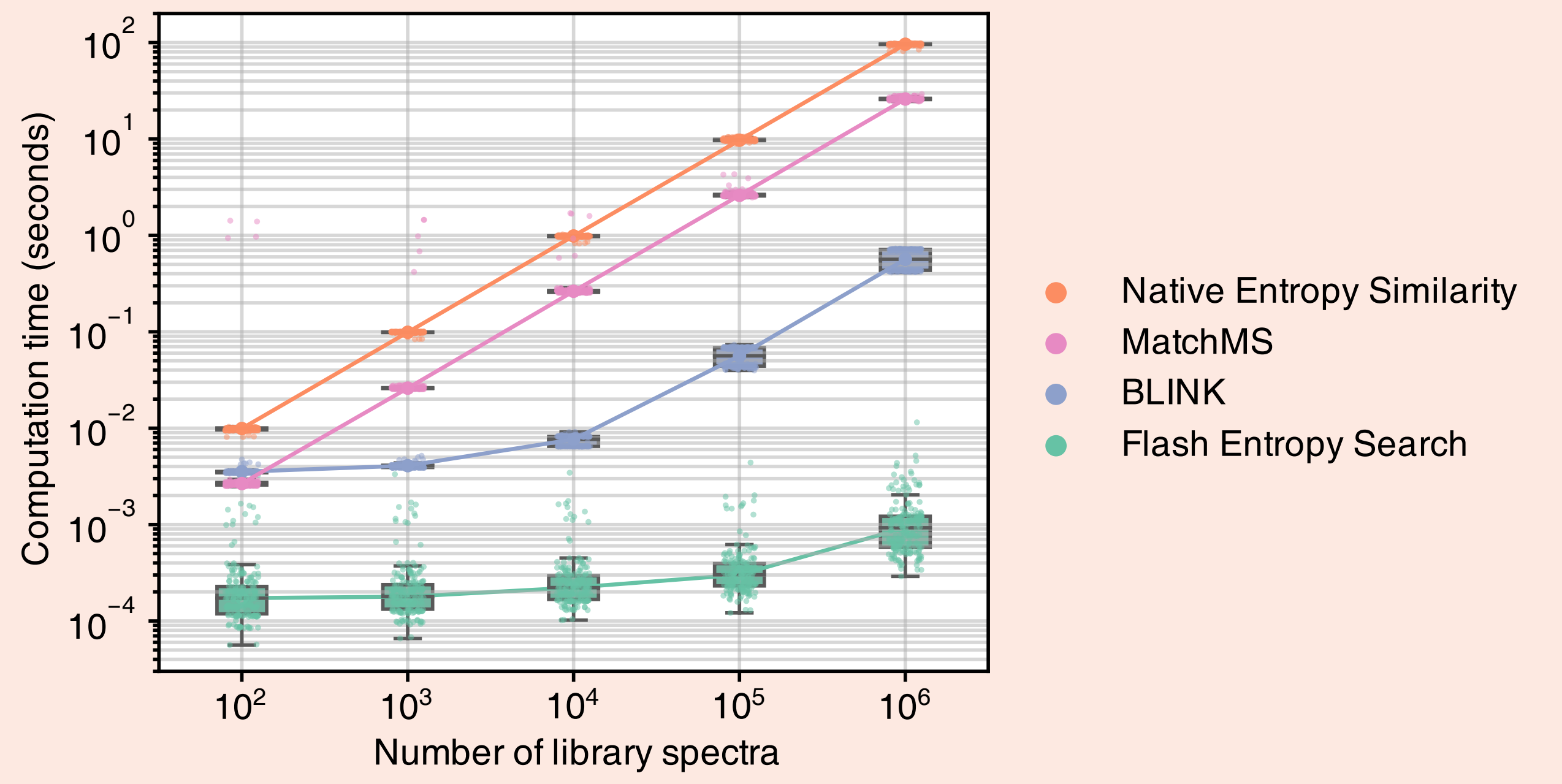


Figure 3. Calculation time to search 100 positive ESI and 100 negative ESI MS/MS spectra against randomly picked samples of the MassBank.us library. Dots represent calculation times per spectrum.

Powerful

Flash entropy search can perform:

✓ Identity search

✓ Open search

✓ Neutral loss search

✓ Hybrid search

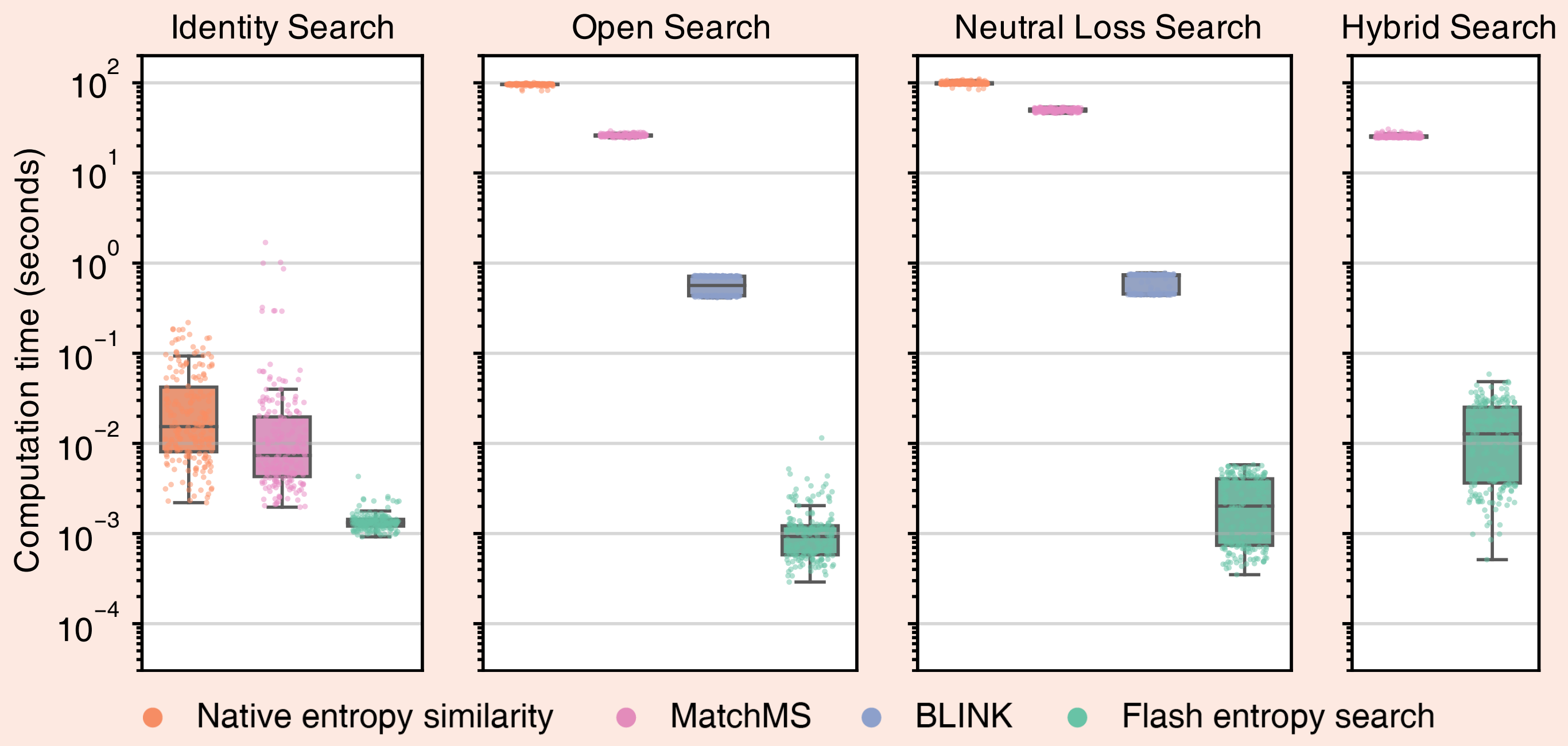


Figure 4. Calculation times to perform identity, open, neutral loss and hybrid searches for 100 positive ESI and 100 negative ESI spectra against 1,000,000 MassBank.us spectra with different algorithms.

Accurate

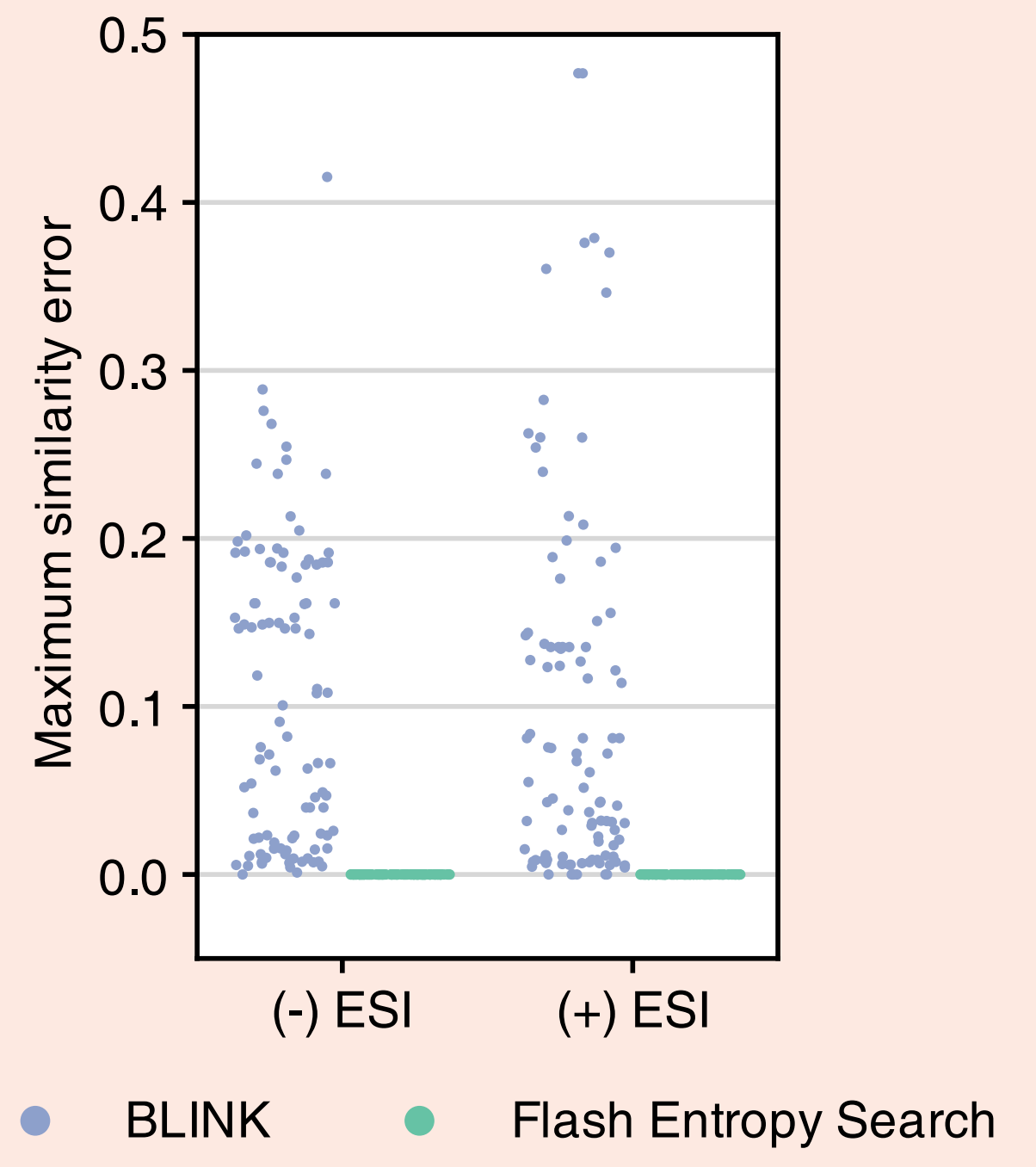


Figure 5. Accuracy of MS/MS similarity results comparing Flash entropy to regular entropy searches and BLINK to MatchMS dot-product scores.

Flexible

Flash entropy search algorithm is very flexible, can be easily modified to calculate other spectral similarity.

For example, it can be modified to calculate dot product similarity at similarity performance.

Excellent multi-threads performance

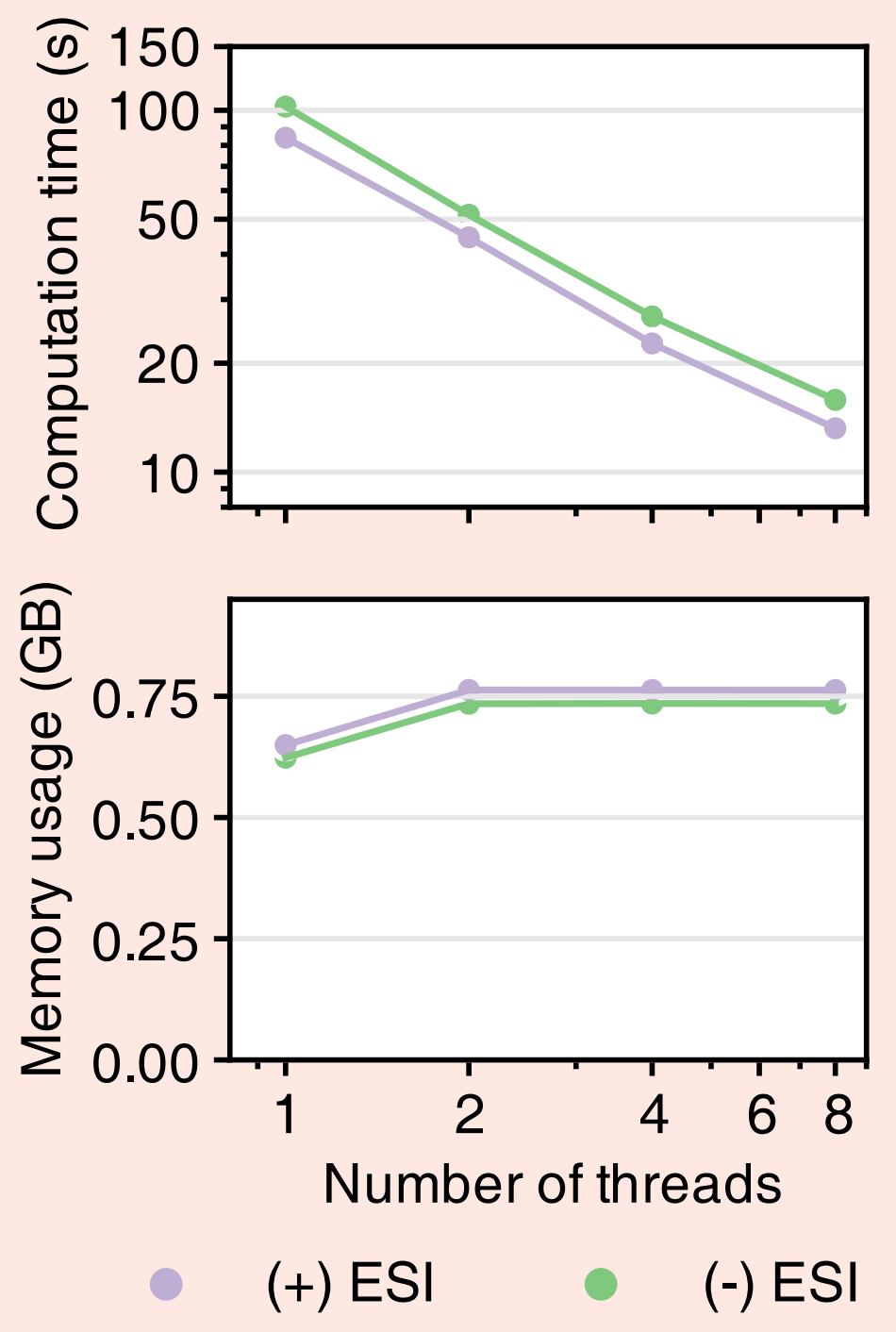


Figure 6. Total computation times and memory usage for conducting an open search of 100,000 spectra against a library of 1,000,000 spectra.

Low memory usage

Boost by GPU

- Flash entropy search need very few memory to search spectra.
- This figure shows using Flash entropy search in a computer with 64 GB to search 738 million spectra in seconds.
- Flash entropy search speed can be boost by GPU.

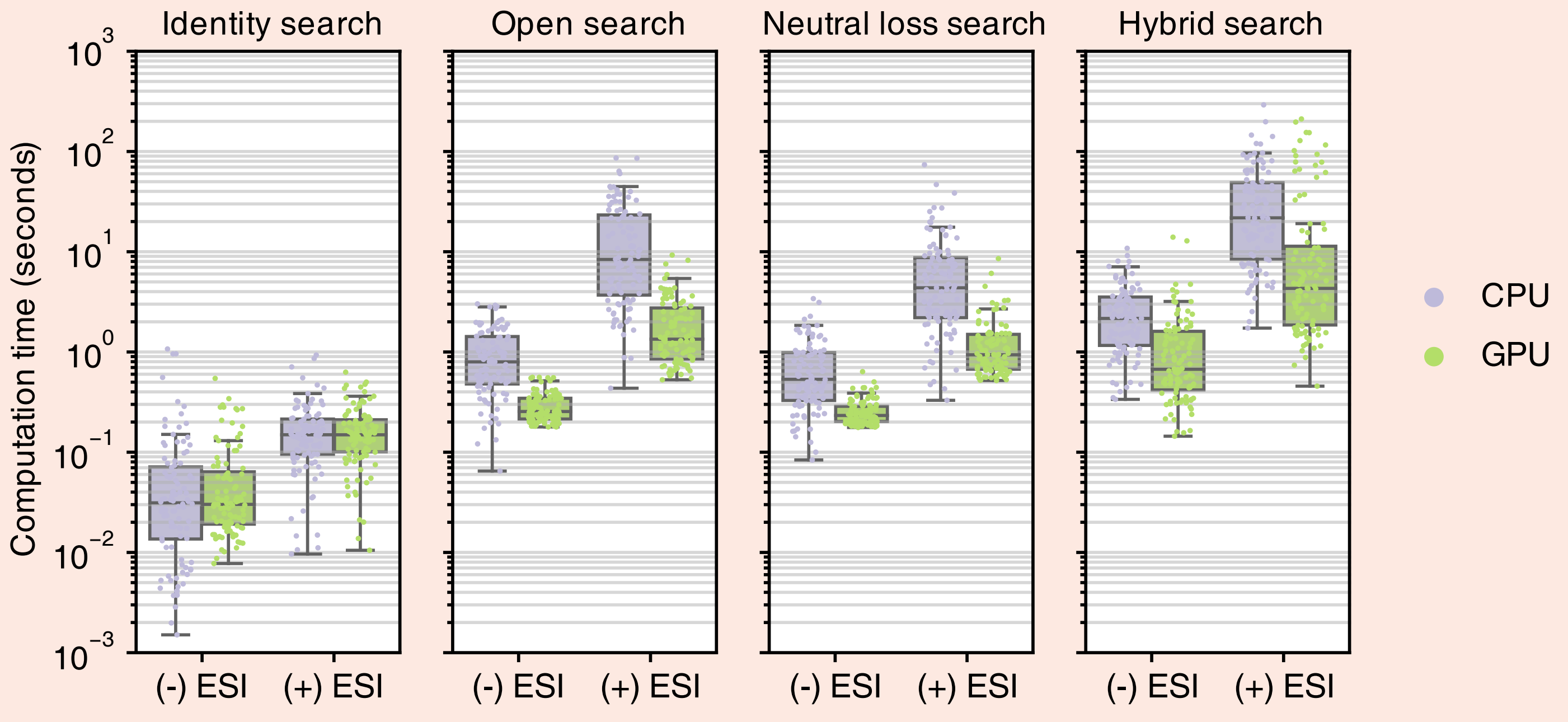


Figure 7: Comparison of computation times when using CPU versus GPU for Flash entropy searches. The 100 negative ESI and positive ESI spectra were searched against 237,185,147 publicly available negative ESI MS/MS spectra and 701,996,947 positive MS/MS spectra.

Conclusion

We developed, implemented and evaluated Flash entropy to calculate similarity-matching of millions of accurate mass MS/MS spectra within less than 10 ms (or a billion spectra in <2 s), using classic low-memory personal computers. Flash entropy presents ultrafast computing on a big-data scale to every laboratory. It extends similarity-matching from simple identity searches to include open, neutral loss and hybrid searches. This method has five benefits over alternative approaches: (1) It greatly improves computation efficiency when comparing large spectral libraries; (2) it does not require binning of the product ions and does not alter the accuracy of similarity results; (3) it can be run in parallel using multiple cores with minimal overhead; (4) it retains high performance when analyzing spectral libraries that are too large to be entirely loaded into the memory; and (5) its speed can be boosted by GPUs.

This study was funded by National Institutes of Health grants U2C ES030158 and R03 OD034497.