Supporting information to "Fast open modification spectral library searching through approximate nearest neighbor indexing"

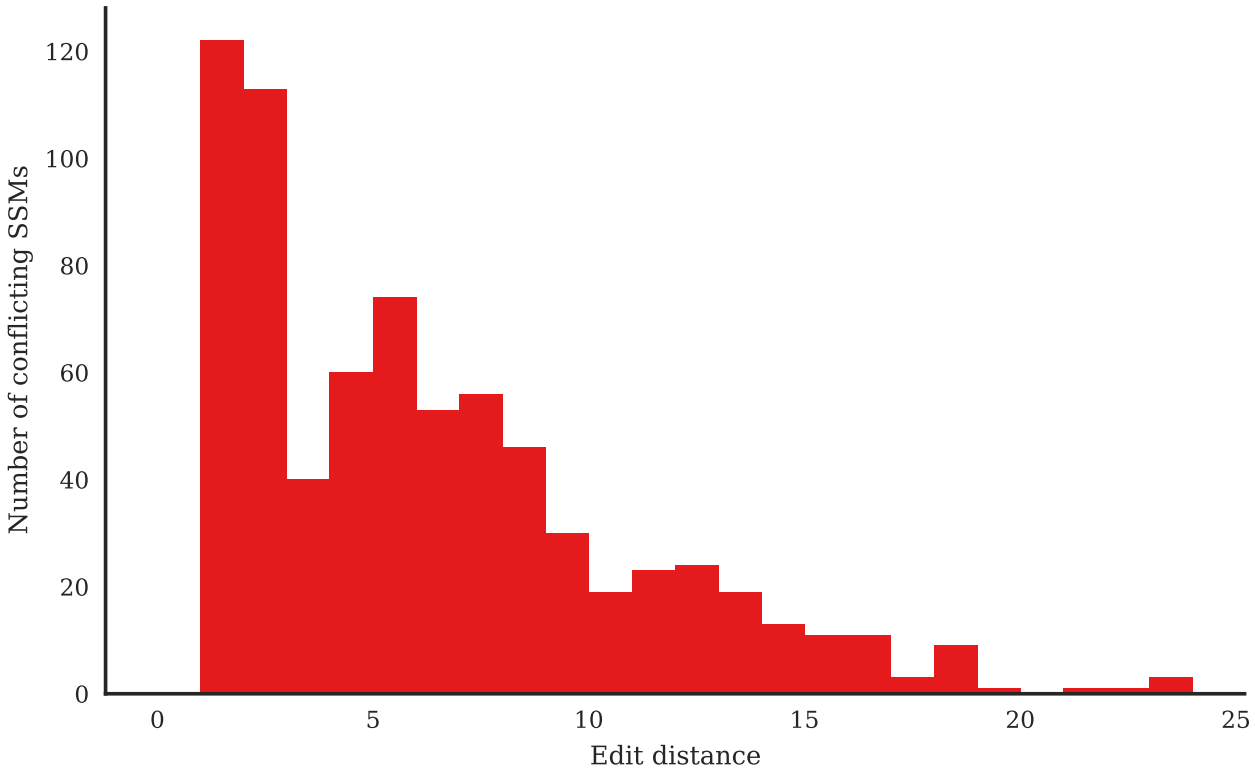
Wout Bittremieux[1,2,3], **Pieter Meysman**[1,2], **William Stafford Noble**[3,4], **Kris Laukens**[1,2]

[1]Department of Mathematics and Computer Science, University of Antwerp, 2020 Antwerp, Belgium; [2]Biomedical Informatics Network Antwerpen (biomina), 2020 Antwerp, Belgium; [3]Department of Genome Sciences, University of Washington, Seattle WA 98195, USA; [4]Department of Computer Science and Engineering, University of Washington, Seattle WA 98195, USA

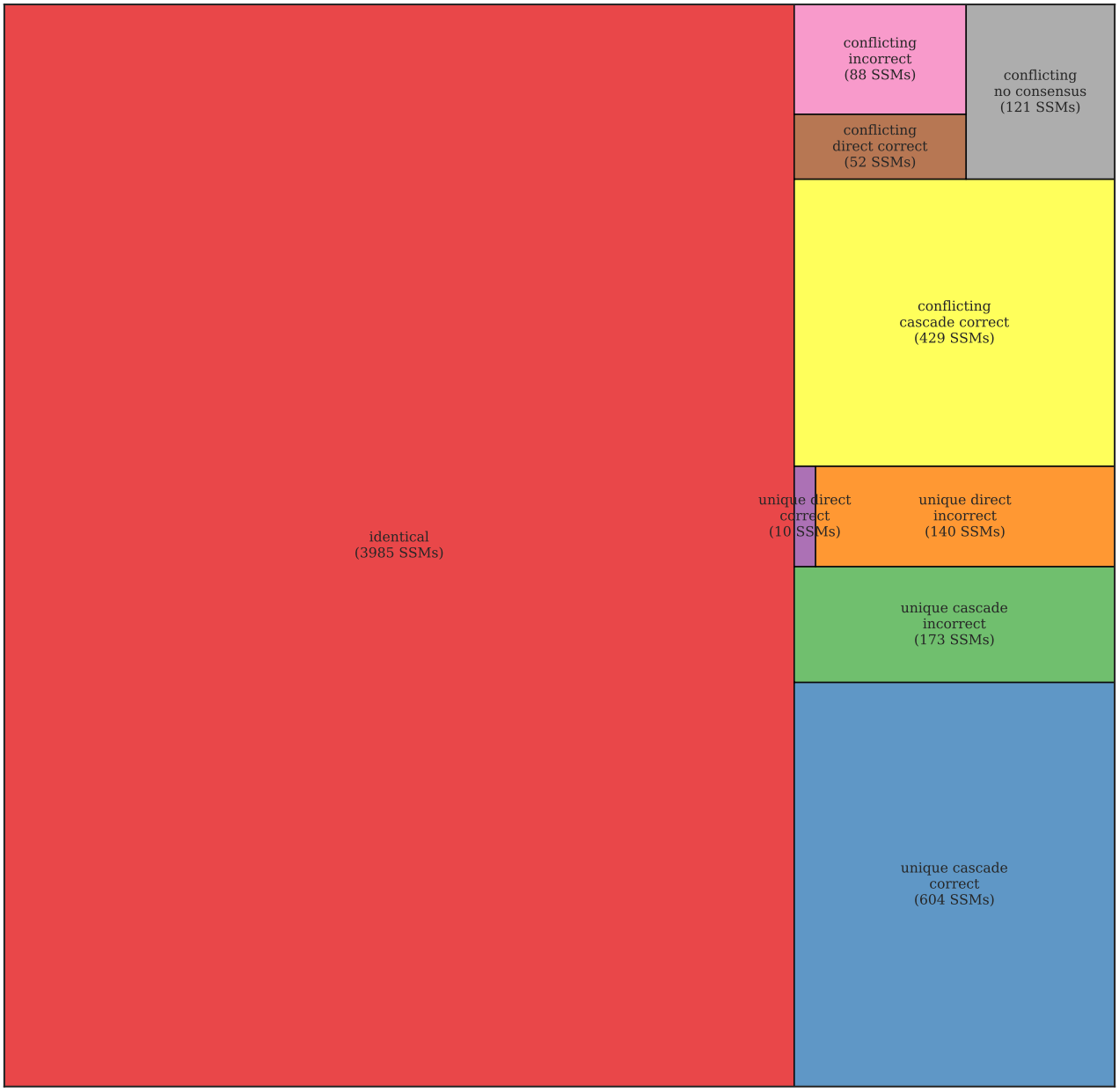
Corresponding author: wout.bittremieux@uantwerpen.be, kris.laukens@uantwerpen.be

List of Figures

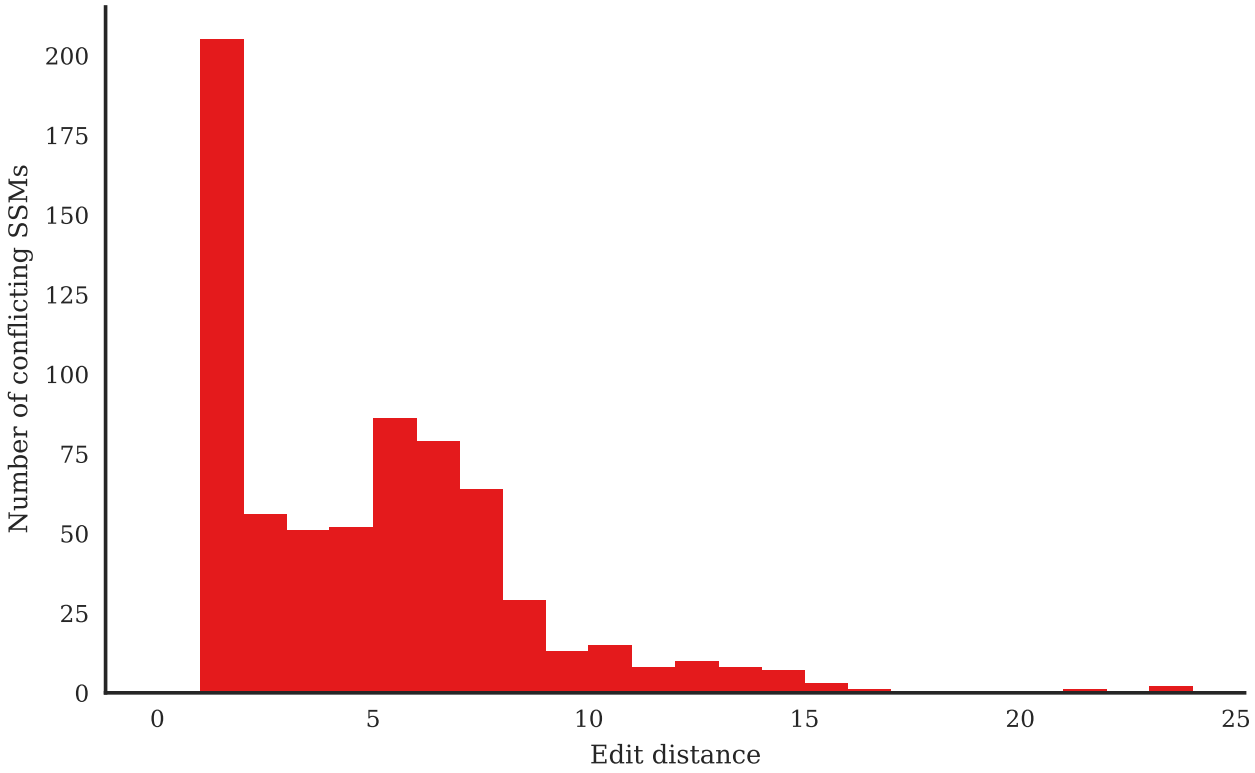
List of Tables



Supplementary Figure S1: Peptide sequence similarity for the conflicting SSMs between ANN-SoLo and the iPRG2012 consensus results. The sequence similarity for each SSM is quantified by the edit distance between the peptide sequence assigned by ANN-SoLo and the peptide sequence from the iPRG2012 consensus results.



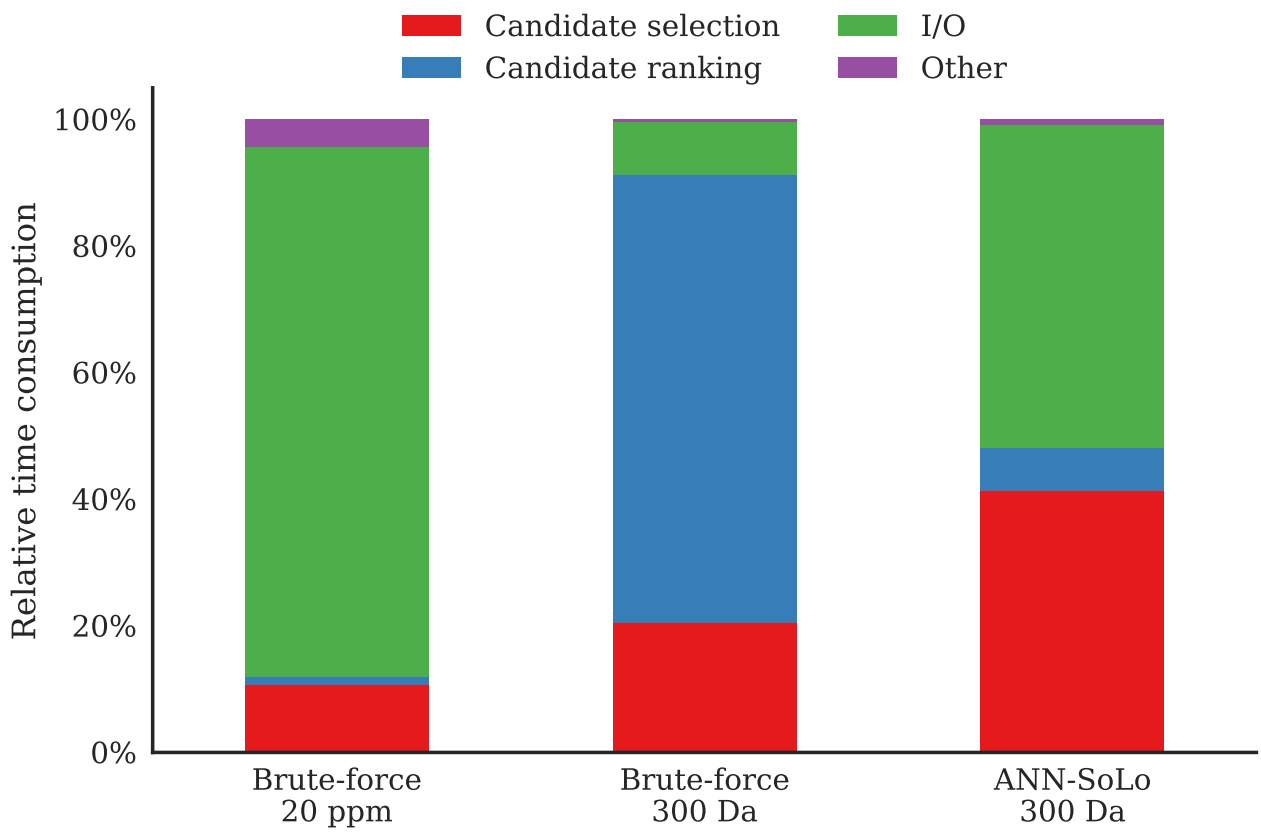
Supplementary Figure S2: Identification comparison between a cascade open search and a direct open search for the iPRG2012 data set. Based on a comparison to the iPRG2012 consensus results the cascade search identifies more correct peptides than the direct open search, both for the SSMs where both searches provide a conflicting peptide assignment and for the non-overlapping SSMs.



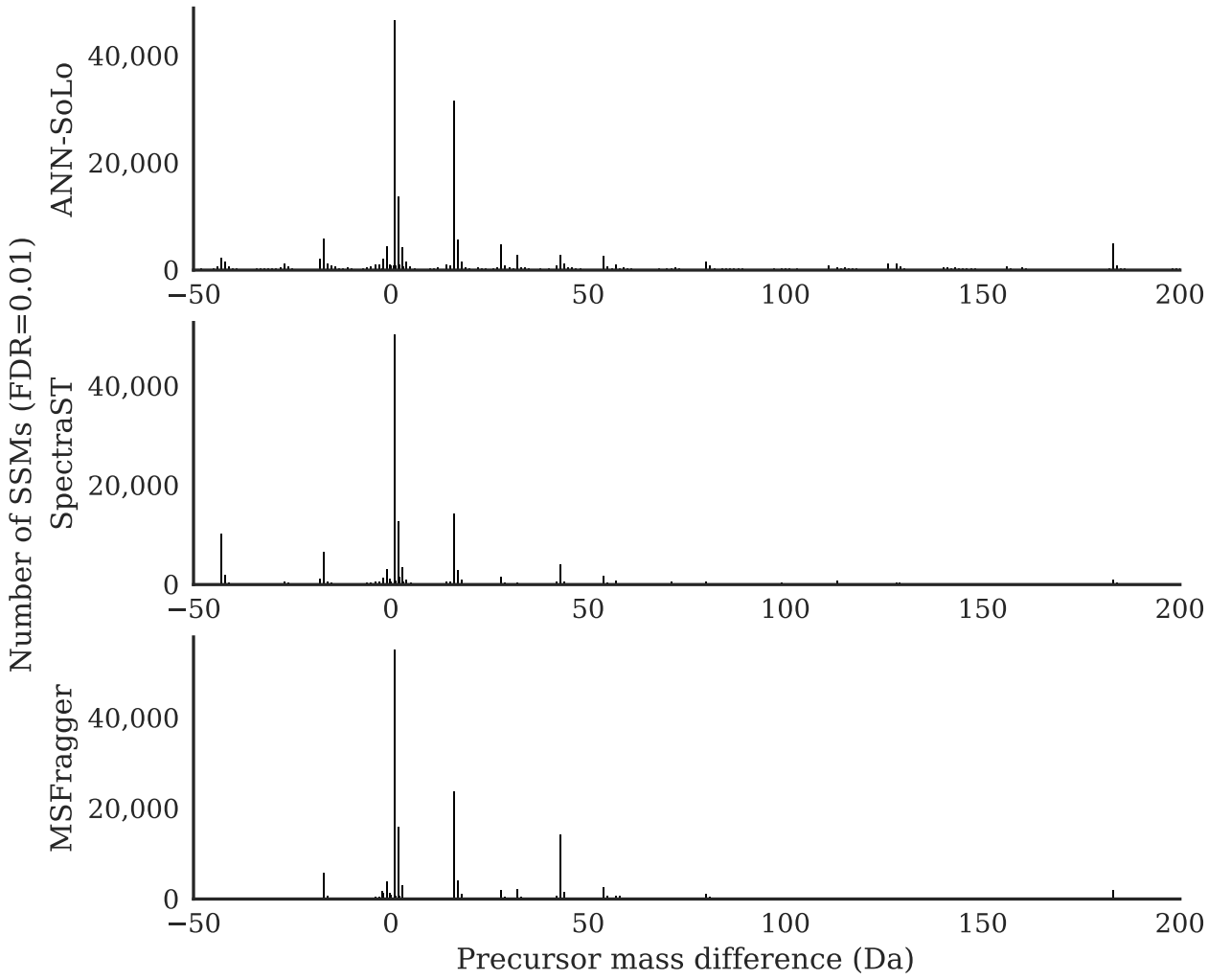
Supplementary Figure S3: Peptide sequence similarity for the conflicting SSMs between a cascade open search and a direct open search for the iPRG2012 data set. The sequence similarity for each SSM is quantified by the edit distance between the peptide sequence assigned by the cascade open search and the peptide sequence assigned by the direct open search.

Search mode	num_trees	Build time (min)	Index size (GB)	search_k	# SSMs	Search time (min)
Brute-force 20 ppm					4141	9.0
Brute-force 300 Da					6019	269.7
ANN-SoLo 300 Da	100	46.1	8.00	20 000	5488	25.9
				40 000	5605	26.7
				100 000	5677	29.9
				200 000	5786	33.5
				400 000	5884	40.1
ANN-SoLo 300 Da	200	68.5	9.83	20 000	5247	25.8
				40 000	5556	26.5
				100 000	5635	29.9
				200 000	5746	33.8
				400 000	5832	39.4
ANN-SoLo 300 Da	500	124.5	15.32	20 000	5474	26.3
				40 000	5634	27.2
				100 000	5701	30.5
				200 000	5745	33.4
				400 000	5839	38.6
ANN-SoLo 300 Da	1000	224.5	24.48	20 000	5423	26.6
				40 000	5569	28.1
				100 000	5703	30.5
				200 000	5794	33.9
				400 000	5839	38.9

Supplementary Table S1: ANN-SoLo index properties and search performance for various num_trees and search_k hyperparameter combinations for the iPRG2012 data set. Timing results were obtained on a single-core Intel Xeon E5-2680 v2 processor. Index build times include the time required to read the entire spectral library into memory and process it prior to index construction, which was around 27 minutes for the described spectral library. The reported ANN index size is the total combined size of all index files; individual files are smaller as separate files are used for different precursor charges.



Supplementary Figure S4: Profiling shows how much time was spent in each part of the code for various searches of the iPRG2012 data set. During a brute-force open search the majority of time is spent during the candidate ranking step, while ANN indexing helps to select only a limited number of candidates and minimize the time required to rank these candidates. Note that relative time consumptions are reported: using an ANN index results in a significant speedup, as shown previously. Correspondingly, for example, although the relative I/O time consumption is higher when using an ANN index than in the brute-force case, the absolute I/O time consumption is lower.



Supplementary Figure S5: Precursor mass differences for ANN-SoLo, SpectraST, and MSFragger for the HEK293 data set. Only non-zero precursor mass differences are shown, whereas the majority of SSMs corresponds to unmodified peptides with a zero precursor mass difference.