Improved open modification searching via unified spectral search with predicted libraries and enhanced vector representations in ANN-SoLo

Issar Arab[1], Kris Laukens[1,2], Wout Bittremieux[1,*]
1 Department of Computer Science, University of Antwerp, 2020 Antwerp, Belgium
2 Biomedical Informatics Network Antwerpen (biomina), 2020 Antwerp, Belgium
* wout.bittremieux@uantwerpen.be

Abstract

Summary: The primary computational challenge in mass spectrometry-based proteomics is determining the peptide sequence responsible for generating each measured tandem mass spectrum. This task is traditionally addressed through sequence database searching, as well as alternative approaches such as spectral library searching. ANN-SoLo is a powerful spectral library search engine optimized for open modification searching, enabling the detection of peptides carrying any post-translational modification. Here, we present an enhanced version of ANN-SoLo that combines strengths of both spectral library searching and sequence database searching, by integrating with Prosit to generate predicted spectral libraries from protein sequence databases. Additionally, it provides functionality to generate decoys at both the spectrum and the peptide level, introduces an optimized internal file structure for large-scale analytics, and improves search accuracy by incorporating neutral loss information into spectrum vector representations. These advancements collectively address challenges associated with missing spectral libraries and enhance peptide identification in large-scale and complex proteomics workflows.

Availability and Implementation: ANN-SoLo is available as open source under the permissive Apache 2.0 license on GitHub at https://github.com/bittremieux-lab/ANN-SoLo.

1. Introduction

Tandem mass spectrometry (MS/MS)-based proteomics is a powerful analytical approach for identifying and quantifying proteins in complex biological samples. The central challenge in this field is to accurately determine the peptide sequences that give rise to observed MS/MS spectra. This peptide identification task is crucial for understanding protein composition, post-translational modifications (PTMs), and biological function (Shuken 2023). In proteomics, two primary computational strategies are employed to address this challenge: sequence database searching and spectral library searching.

Sequence database searching, the more widely adopted approach, relies on comparing experimental spectra to theoretical spectra derived from a protein sequence database (Eng et al. 2011). This strategy enables the identification of any peptide present in the database, regardless of whether it has been previously observed. However, a key limitation is that theoretical spectra lack accurate fragment ion intensity information, making them only approximate representations of real spectra. This can impact the precision of peptide identifications, particularly in complex samples. Despite this drawback, sequence database searching remains the dominant method due to its broad applicability, as it is not constrained by the need for pre-existing spectral data.

In contrast, spectral library searching matches experimental spectra against a collection of previously acquired reference spectra (Griss 2016, Shao and Lam 2017). This approach improves accuracy by incorporating real fragment ion intensity patterns, leading to more sensitive peptide-spectrum matching. Furthermore, spectral library searching is computationally efficient because it only compares against known, experimentally observed peptides. This reduces the search space and makes it much faster than generating and comparing theoretical spectra for all possible peptides in a full protein database. However, this makes the spectral library searching inherently limited to peptides already present in the library, meaning novel or uncharacterized peptides cannot be found.

Open modification searching (OMS) enhances peptide identification by using a wide precursor mass window, allowing the detection of diverse modifications, including PTMs, amino acid substitutions, and cleavage variants. ANN-SoLo (Bittremieux et al. 2018, Bittremieux et al. 2019, Arab et al. 2023) stands out as a highly efficient spectral library search engine optimized for OMS to detect peptides with any mass shift, enabling the identification of modified variants and overcoming the limitations of traditional closed-search approaches.

In this work, we present an enhanced version of ANN-SoLo that integrates with Prosit (Gessulat et al. 2019) to generate predicted spectral libraries from protein sequence databases, expands support for a wider range of input library and query file formats, handles decoy generation internally, introduces an optimized internal file structure designed for large-scale analytics, and improves search accuracy by incorporating neutral loss information into vector-based spectral representations. These advancements collectively address challenges associated with missing spectral libraries, enhance the accuracy and efficiency of peptide identification across diverse proteomics workflows, and enable efficient searches on large-scale spectral libraries.

2. Results

The latest release of ANN-SoLo (version 0.4.0) introduces several new features and enhancements designed to improve flexibility, processing speed, and overall usability of the software (Figure 1a). In the following sections, we detail the individual new functionalities and their impact on ANN-SoLo's performance and usability. The new features were assessed using the iPRG2012 data set (Chalkley et al. 2014) across three distinct search strategies (https://zenodo.org/records/15164430), with further details provided in Supplementary Section 1. In Supplementary Section 2, we outline ANN-SoLo's data preprocessing and rescoring configuration, covering spectrum filtering and peptide-spectrum match (PSM) rescoring, while Supplementary Section 3 summarizes the software stack, key dependencies, and recent upgrades.

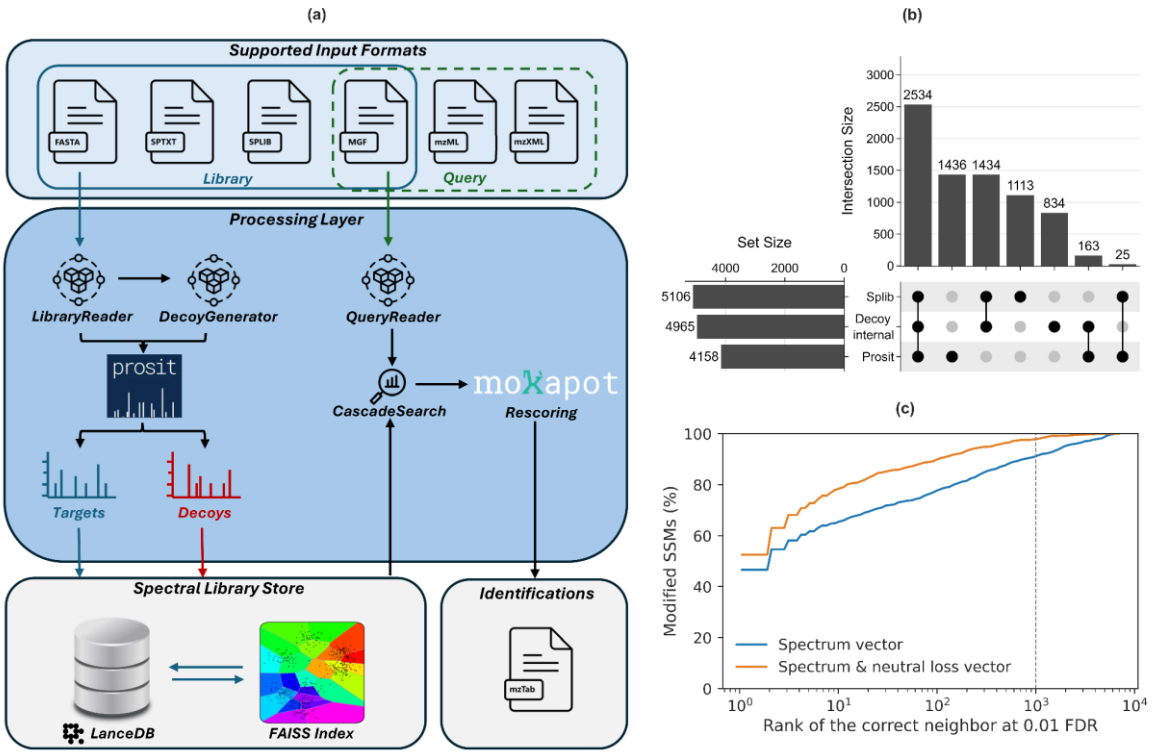

Figure 1. **(a)** High-level architecture diagram illustrating the search process of a query file against a spectral library generated from a FASTA file as input. **(b)** UpSet plot comparing PSM-level search results across three strategies: an splib library with decoys generated via SpectraST, a spectral library with internally generated decoys, and a Prosit-generated library with internal decoy generation. **(c)** Improved ranking of the true PSM during candidate retrieval by incorporating the neutral loss information.

Spectrum Reading From Various Query and Library File Formats

The first improvement is the expansion of query and library file format support. Previously limited to only MGF (Mascot Generic Format), users can now provide query MS/MS input in two additional formats, including mzML (Martens et al. 2011) and mzXML. Moreover, ANN-SoLo now supports spectral libraries in the MGF and the text-based sptxt and binary splib formats used by SpectraST (Lam et al. 2007). Furthermore, it now accommodates FASTA as input files, making it possible to search against comprehensive protein sequence databases, as described further. This variety of file format support ensures that ANN-SoLo can seamlessly interface with a diverse range of input data types, streamlining the workflow for users.

Efficient Internal Storage Format Leveraging LanceDB

The second key enhancement comes from the integration of Lance (Lance 2023) and LanceDB (LanceDB 2024) as internal storage format. Lance is a high-performance columnar data format designed specifically to handle the demands of modern advanced analytics and big data processing. This new internal format accelerates both the insertion of bulk data and querying operations, with the tool now able to handle up to 500,000 spectra in bulk insertions and querying as many as 100,000 spectra per operation. Efficient querying is supported by scalar indexing on two attributes of the spectra: a B-tree index on the identifier, enabling binary search on this column, and a bitmap index on the precursor charge for efficient filtering. This feature enhancement is particularly important for large-scale proteomics studies, where efficient data I/O is crucial as modern MS/MS datasets continue to grow in size.

Decoy Generation Using the Shuffle-and-Reposition Strategy

Previously, generating decoy libraries required the use of external tools, such as SpectraST. This functionality is now integrated directly into ANN-SoLo, based on the shuffle-and-reposition strategy (Lam et al. 2010). First, the peptide sequence is randomly shuffled, while keeping key residues fixed (namely lysine (K), arginine (R), proline (P), and the last residue in the sequence). Second, spectrum_utils (Bittremieux et al. 2023) is used to annotate fragment ions for the target peptide and the annotated peaks are moved to the new *m/z* values according to the decoy sequence. A default edit distance threshold of 70% is applied to guarantee adequate variation between the target and decoy sequences, with up to 10 iterations attempted per target. If no valid decoy is found, decoy generation for that target is discarded. Additionally, all processed peptides, including the generated decoys, are standardized into the ProForma 2.0 format (LeDuc et al. 2022).

The internal decoy generation feature produces outcomes comparable to those using decoys generated with external tools like SpectraST (Figure 1b), demonstrating its reliability and accuracy. Meanwhile, providing this functionality within ANN-SoLo enhances its usability by eliminating the need for external software in this key analysis step, while also still supporting libraries with externally generated decoys. Supplementary Table S1 outlines the configuration parameters for enabling internal decoy generation.

Built-In Spectrum Simulation Using Prosit

Sequence database searching and spectral library searching have traditionally been considered distinct strategies. Sequence databases offer flexibility by allowing identification of any peptide present in the database, even if it has never been observed, while spectral libraries provide higher sensitivity by leveraging experimentally derived fragment ion intensities. However, with recent advances in fragment ion intensity prediction, it is now possible to simulate proteome-scale spectral libraries. Unlike other approaches that use predicted fragment ion intensities only during rescoring, we directly search against predicted spectra, effectively harmonizing sequence database searching and spectral library searching. When a FASTA file is provided as input, protein sequences are cleaved into peptides according to the protease specified in the configuration—by default, trypsin. The resulting peptides are then processed by Prosit (Gessulat et al. 2019), through the Koina interface (Lautenbacher et al. 2024), to generate predicted spectra, which are stored in the internal spectral library format for future search routines. Notably, Prosit has also been used in other workflows such as Scribe (Searle et al., 2023), where it was employed to predict both fragmentation and retention times for all peptides derived from a FASTA database. A full list of configurable parameters for FASTA file processing can be found in Supplementary Table S1.

To evaluate this feature, we analyzed the iPRG2012 dataset using default settings. As detailed in Supplementary Section 1, spectra with precursor charges 2 and 3 were generated using trypsin cleavage, resulting in a spectral library containing ~14.5 million spectra. The search identified peptides for 4,158 query spectra at 1% FDR (Figure 1b). Overall, the splib spectral library yielded more results, due to not only canonical peaks but additional peaks used for more sensitive matching. However, as database searching considers a larger search space, it can identify certain peptides that might be entirely missed by library searching. For example, this is evident for MS/MS spectra with precursor charge 3, which are not comprehensively covered by the spectral library, and thus could be uniquely identified using database searching (Supplementary Figure S1). Among the newly identified PSMs found exclusively using the Prosit-generated library, approximately 92% of peptides were absent from the splib target library (Supplementary Figure S2), demonstrating the increased coverage and unique spectral information provided by Prosit-generated spectra.

Optimized Spectrum-to-Vector Conversion Using Neutral Loss Integration

ANN-SoLo employs a two-pass strategy for spectrum annotation. First, it uses approximate nearest neighbor searching to retrieve candidates based on the vector-based similarity of hashed MS/MS spectra (Bittremieux et al. 2018, Bittremieux et al. 2019). Next, it re-ranks these candidates using native spectrum similarity, with modified cosine as the default function.

To build a nearest neighbor index to efficiently select candidates from the spectral library, spectra are vectorized to represent them as points in a multidimensional space. The following two-step procedure is used to convert a high-resolution MS/MS spectrum to a vector: (1) we convert the spectrum to a sparse vector using small mass bins to tightly capture fragment masses. (2) We convert the sparse, high-dimensional vector to a lower-dimensional vector by using a hash function to map the mass bins to a limited number of hash bins (Bittremieux et al. 2019). In mathematical notation, let $h: \mathbb{N} \rightarrow \{1, \ldots, m\}$ be a random hash function. Then h can be used to convert a vector $x = \{x_1, \ldots, x_n\}$ to a vector $x' = \{x'_1, \ldots, x'_m\}$, with $m \ll n$:

$$x'_i = \sum_{j:h(j)=i} x_j \qquad (1)$$

Neutral loss refers to the mass difference between a precursor ion and its fragment ions in MS/MS spectra, representing the loss of molecular substructures during fragmentation. During neutral loss matching, MS/MS spectra are mirrored at their precursor mass by calculating the distances from each fragment ion peak to the precursor mass, describing the neutral losses (Bittremieux et al. 2022). In this new version of the software, we enhance the vector representation by integrating the neutral loss information. The updated vectorization process follows these four steps: (1) the spectrum is projected as before, generating a fixed-size vector of length m; (2) the neutral loss spectrum is computed; (3) this neutral loss spectrum undergoes the same transformation using the hashing function 1, producing another fixed-size vector of length m; (4) both vectors are concatenated, resulting in a final vector representation of size $2m$ (Supplementary Figure S3).

With the integration of neutral loss information, ANN-SoLo further refines its candidate selection, aligning nearest neighbor scoring more closely with the final spectrum-based modified cosine scores. This approach allows ANN-SoLo to reduce the number of candidate peptides it needs to evaluate while still preserving most identifications (Figure 1c). As a result, the software achieves faster search speeds without compromising result quality, making it more efficient for large-scale proteomics analyses.

3. Conclusion

The enhancements introduced in ANN-SoLo (version 0.4.0) bring significant improvements in both functionality and performance. The diverse query and library file support enables a greater variety of data formats to be handled, providing flexibility in the analysis of proteomic data. The adoption of LanceDB as the internal storage format has streamlined the pipeline, enhancing speed and efficiency while laying the groundwork for future integrations with other data formats and facilitating the processing of large datasets. The integration of an internal decoy generation module has simplified the workflow, eliminating the need for external tools and ensuring reliable and efficient decoy creation. The inclusion of Prosit-based spectrum predictions further enhances the tool's capabilities, providing unique peptide identifications that were previously not captured by traditional spectral libraries. The integration of neutral loss information into candidate retrieval has also significantly improved search accuracy and speed, optimizing candidate peptide ranking and reducing computational overhead.

The introduced new features along with the already existing ones, namely the approximate nearest neighbor-based open search and the rescoring module, have made ANN-SoLo a highly versatile method. These features not only increase ANN-SoLo's speed and scalability but also improve the accuracy, reliability, and ease of use for proteomics researchers. As proteomics continues to advance and datasets become increasingly large and complex, the new features in ANN-SoLo ensure that it remains a powerful, efficient, and versatile tool for researchers. Looking forward, future developments could further expand the tool's capabilities, including the integration of additional data formats and continued refinement of its algorithms to enhance both accuracy and speed.

References

Shuken SR. An introduction to mass spectrometry-based proteomics. Journal of proteome research. 2023 Jun 1;22(7):2151-71.

Eng JK, Searle BC, Clauser KR, Tabb DL. A face in the crowd: recognizing peptides through database search. Molecular & Cellular Proteomics. 2011 Nov 1;10(11).

Griss J. Spectral library searching in proteomics. Proteomics. 2016 Mar;16(5):729-40.

Shao W, Lam H. Tandem mass spectral libraries of peptides and their roles in proteomics research. Mass spectrometry reviews. 2017 Sep;36(5):634-48.

Bittremieux W, Meysman P, Noble WS, Laukens K. Fast open modification spectral library searching through approximate nearest neighbor indexing. Journal of proteome research. 2018 Sep 5;17(10):3463-74.

Bittremieux W, Laukens K, Noble WS. Extremely fast and accurate open modification spectral library searching of high-resolution mass spectra using feature hashing and graphics processing units. Journal of proteome research. 2019 Aug 26;18(10):3792-9.

Arab I, Fondrie WE, Laukens K, Bittremieux W. Semisupervised machine learning for sensitive open modification spectral library searching. Journal of proteome research. 2023 Jan 23;22(2):585-93.

Gessulat S, Schmidt T, Zolg DP, Samaras P, Schnatbaum K, Zerweck J, Knaute T, Rechenberger J, Delanghe B, Huhmer A, Reimer U. Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. Nature methods. 2019 Jun;16(6):509-18.

Searle BC, Shannon AE, Wilburn DB. Scribe: next generation library searching for DDA experiments. Journal of Proteome Research. 2023 Jan 25;22(2):482-90.

Lautenbacher L, Yang KL, Kockmann T, Panse C, Chambers M, Kahl E, Yu F, Gabriel W, Bold D, Schmidt T, Li K. Koina: Democratizing machine learning for proteomics research. bioRxiv. 2024 Jun 3.

Chalkley RJ, Bandeira N, Chambers MC, Clauser KR, Cottrell JS, Deutsch EW, Kapp EA, Lam HH, McDonald WH, Neubert TA, Sun RX. Proteome informatics research group (iPRG) _2012: a study on detecting modified peptides in a complex mixture. Molecular & Cellular Proteomics. 2014 Jan 1;13(1):360-71.

Martens L, Chambers M, Sturm M, Kessner D, Levander F, Shofstahl J, Tang WH, Römpp A, Neumann S, Pizarro AD, Montecchi-Palazzi L. mzML—a community standard for mass spectrometry data. Molecular & Cellular Proteomics. 2011 Jan 1;10(1).

Lance. Lance: modern columnar data format for ML [Software]. 2023. Available from: https://lancedb.github.io/lance/

lancedb/lancedb. Lance DB [Software]. 2024 Dec 1 [cited 2024 Dec 1]. Available from: https://github.com/lancedb/lancedb

Lam H, Deutsch EW, Aebersold R. Artificial decoy spectral libraries for false discovery rate estimation in spectral library searching in proteomics. Journal of proteome research. 2010 Jan 4;9(1):605-10.

Bittremieux W, Levitsky L, Pilz M, Sachsenberg T, Huber F, Wang M, Dorrestein PC. Unified and standardized mass spectrometry data processing in Python using spectrum_utils. Journal of proteome research. 2023 Jan 23;22(2):625-31.

LeDuc RD, Deutsch EW, Binz PA, Fellers RT, Cesnik AJ, Klein JA, Van Den Bossche T, Gabriels R, Yalavarthi A, Perez-Riverol Y, Carver J. Proteomics Standards Initiative's ProForma 2.0: Unifying the encoding of proteoforms and peptidoforms. Journal of proteome research. 2022 Mar 15;21(4):1189-95.

Bittremieux W, Schmid R, Huber F, van der Hooft JJ, Wang M, Dorrestein PC. Comparison of cosine, modified cosine, and neutral loss based spectrum alignment for discovery of structurally related molecules. Journal of the American Society for Mass Spectrometry. 2022 Aug 12;33(9):1733-44.

Lam H, Deutsch EW, Eddes JS, Eng JK, King N, Stein SE, Aebersold R. Development and validation of a spectral library searching method for peptide identification from MS/MS. Proteomics. 2007 Mar;7(5):655-67.