Comparison of Cosine, Modified Cosine, and Neutral Loss Based Spectral Alignment For Discovery of Structurally Related Molecules

Wout Bittremieux[1,2,#], Robin Schmid[1,2,#], Florian Huber[3], Justin JJ van der Hooft[4,5], Mingxun Wang[1,2], Pieter C Dorrestein[1,2*]

1. Collaborative Mass Spectrometry Innovation Center, University of California San Diego, La Jolla, CA 92093, USA
2. Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, La Jolla CA 92093, USA
3. Centre for Digitalization and Digitality, University of Applied Sciences, 40476 Düsseldorf, Germany
4. Bioinformatics Group, Wageningen University, 6708PB Wageningen, the Netherlands
5. Department of Biochemistry, University of Johannesburg, Auckland Park, Johannesburg 2006, South Africa

[#] Equal contribution.
* Corresponding author: pdorrestein@health.ucsd.edu

Abstract

Spectral alignment of tandem mass spectrometry (MS/MS) data using the modified cosine similarity and subsequent visualization as molecular networks have been demonstrated to be a useful strategy to discover analogs of molecules from untargeted MS/MS-based metabolomics experiments. Recently, a neutral loss matching approach has been introduced as an alternative to MS/MS-based molecular networking, with an implied performance advantage in finding analogs that cannot be discovered using existing MS/MS spectral alignment strategies. To comprehensively evaluate the scoring properties of neutral loss matching, the cosine similarity, and the modified cosine similarity, similarity measures of 955,228 peptide MS/MS spectrum pairs and 10 million small molecule MS/MS spectrum pairs were compared. This comparative analysis revealed that the modified cosine similarity outperformed neutral loss matching and the cosine similarity in all cases. The data further indicated that the performance of neutral loss matching depends on both the location and type of the modification. For full transparency and reproducibility, and reusability for any future comparisons of newly developed mass spectral alignment approaches, all code and benchmark data that was used in this work are freely accessible to the community.

Introduction

With the growing availability of free, open, and proprietary tandem mass spectrometry (MS/MS) libraries, MS/MS-based spectrum alignment is routinely used for metabolite annotation and organization of mass spectral datasets.[1–4] For MS/MS spectrum comparison, the cosine

similarity is the most widely used approach to match near-identical spectra with each other.[5,6] Adaptations hereof, such as the modified cosine similarity,[1,7] can be used to discover non-identical but related spectra using analog searching and molecular networking of small molecules.[8,9] Similar approaches have also been used in proteomics to perform open modification searching for the unbiased detection of peptides that contain post-translational modifications.[10–13] Additionally, spurred by the emergence of increasingly powerful computational and machine learning techniques, several novel similarity scores have been proposed.[14–18]

Recently, neutral loss-based spectrum alignment was introduced in the METLIN analysis ecosystem[19] as a strategy to match MS/MS spectra of analog molecules.[20] 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. Effectively, the neutral loss approach recalibrates spectra with their precursor masses as the origin, and a mirror database of neutral loss spectra was created in METLIN.[19] The neutral loss spectra can then be used to find related spectra of modified analog molecules. Based on representative examples, Aisporna *et al.*[20] have demonstrated that neutral loss matching can be particularly effective at finding MS/MS spectrum pairs that other approaches, such as spectrum alignment using the cosine and modified cosine similarities, could not find. This benefit was explained by the increase in spectrum usage (i.e. more explained intensity captured by matching fragment ions) in the assignment of related matches. Nevertheless, although a few select MS/MS spectra were chosen to compare different spectrum similarity measures, the difference in spectrum matching performance was not systematically quantified.[20]

We hypothesized that the modified cosine similarity, which forms the foundation of molecular networking in the Global Natural Products Social Molecular Networking (GNPS) system,[7] should be able to discover all MS/MS spectrum pairs that can be found by neutral loss matching, as it considers both directly matching peaks and matching peaks that are shifted by the pairwise precursor mass difference. We, therefore, set out to systematically compare the performance of neutral loss matching, cosine similarity, and modified cosine similarity on a large collection of spectrum pairs of modified peptides and small molecules. As the data of the original study[20] was not available due to the access restrictions of the METLIN database,[19] we instead used a large collection of 955,228 and 10 million MS/MS spectrum pairs for peptides and small molecules, respectively, which were derived from the MassIVE-KB[21] and GNPS resources.[7]

Results

We evaluated neutral loss matching, cosine similarity, and modified cosine similarity in matching MS/MS spectra of non-identical yet structurally related molecules (**Figure 1**). The cosine similarity is a common strategy to compare MS/MS spectra to each other by matching fragment ions in both spectra with identical *m*/*z* values (while accounting for a relevant fragment mass tolerance based on the data acquisition settings). The modified cosine similarity is an extension of the cosine similarity which not only matches fragment ions with identical *m*/*z* values, but also considers fragment ions across both spectra with *m*/*z* values that are shifted by the precursor mass difference of the spectrum pair under consideration. Finally, neutral loss matching first

transforms MS/MS spectra into neutral loss spectra by mirroring the fragment ion *m*/*z* values at the precursor mass, after which the neutral loss peaks with identical Δ*m*/*z* values are matched to each other. Although MS/MS spectrum preprocessing and filtering,[22] for example by removing noise peaks and rescaling fragment intensities, can be useful to obtain optimally discriminative spectrum similarities, here we did not apply such data transformations to focus on a performance evaluation of the different mass spectrum similarity measures and avoid that the results were influenced by extraneous data processing.

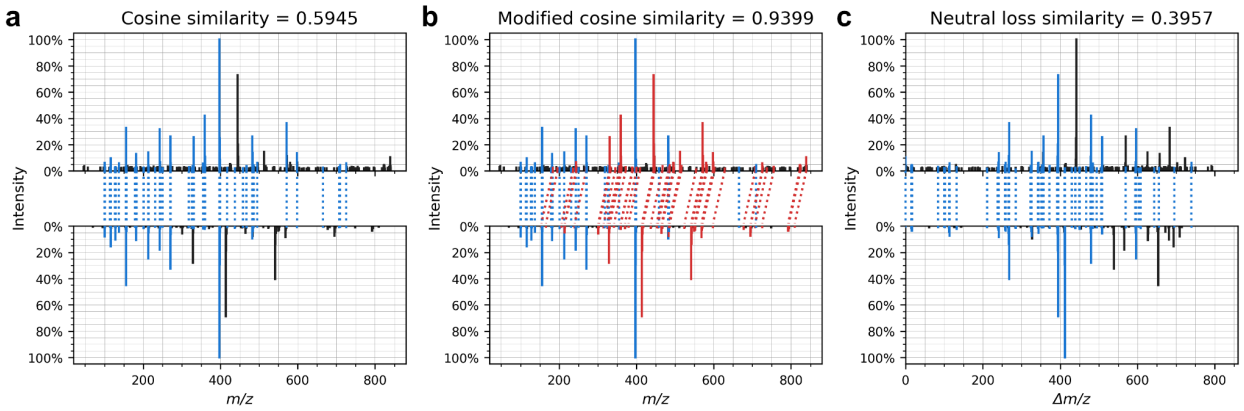

Figure 1. Spectrum similarity measures for the natural product apratoxin A (top; precursor *m*/*z* = 840.497**) and a newly discovered apratoxin analog (bottom**; precursor *m*/*z* = 810.487**).**[23] Illustration of the cosine similarity **(a)**, modified cosine similarity **(b)**, and neutral loss matching **(c**, visualized as neutral loss spectra**)** to compare MS/MS spectra to each other. Directly matching fragments with near-identical *m*/*z* or Δ*m*/*z* values in the original MS/MS spectra and the neutral loss spectra, respectively, are indicated in blue. Additional matching fragments that are offset by the pairwise precursor *m*/*z* difference, considered by the modified cosine similarity, are indicated in red. Unmatched fragments are indicated in black.

Three datasets with heterogeneous mass spectral properties were used to compare the different spectrum similarity measures. Because large-scale ground truth benchmark datasets of related small molecule spectrum pairs with known structural modifications currently do not exist yet, the first dataset we used consisted of 955,228 known peptide MS/MS spectrum pairs derived from the MassIVE-KB resource,[21] where peptide pairs differ by a single modification (post-translational modification or amino acid substitution). Second, mass spectrum pairs were derived from 495,600 small molecule reference MS/MS spectra contained in the GNPS community mass spectral libraries.[7] Because over 1.5 billion pairwise comparisons (precursor mass difference between 1–200 Da) can be performed using these spectra, a random subset of 10 million spectrum pairs was used for computational efficiency. Third, 340,637 bile acid MS/MS spectrum pairs from the GNPS community spectral libraries were used to evaluate the influence of specific modification properties, such as addition or loss of oxygen, conjugation of the bile acids with amino acids (as amino acid amidates), and substitution of conjugated amino acids attached to the bile acid core.

Evaluation of the different spectrum similarity measures on the peptide data indicates that the modified cosine similarity strictly outperforms both the cosine similarity and neutral loss

matching (**Figure 2a**), while neutral loss matching resulted in a higher score than the cosine similarity in only 8.8% of the peptide pairs. Spectrum usage (in terms of the number of explained fragment ions and total explained intensity) can be used to assess the performance of spectrum similarity measures (**Figure 2b**).[20] In general, neutral loss matching resulted in very low spectrum usage for peptide mass spectra, with a median explained intensity of 17.5%. In contrast, the median explained intensity for the cosine similarity and modified cosine similarity were 56.6% and 70.9%, respectively. Taking into account the modification position relative to the linear peptide sequences (**Figure 2c**), neutral loss matching performed better with modifications located closer to the C-terminal end. In this case, a maximal number of y-ions, which are the dominant signals in peptide mass spectra, include the modification and can be matched using the neutral loss strategy. In contrast, the cosine similarity performed best when the modification was located close, but not entirely at the peptide N-termini. We hypothesize that in this case the cosine similarity can match a large number of y-ions as well as the initial b-ions, whereas when the modification occurs at the N-terminus, no b-ions contribute to the cosine similarity while higher index y-ions often do not occur and thus do not provide further benefit. Finally, the modified cosine similarity achieves high scores across a range of modification positions (except for N-terminal modifications), which indicates its beneficial performance in capturing both directly matching fragment ions and fragment ions that are shifted due to the mass difference induced by the modification.

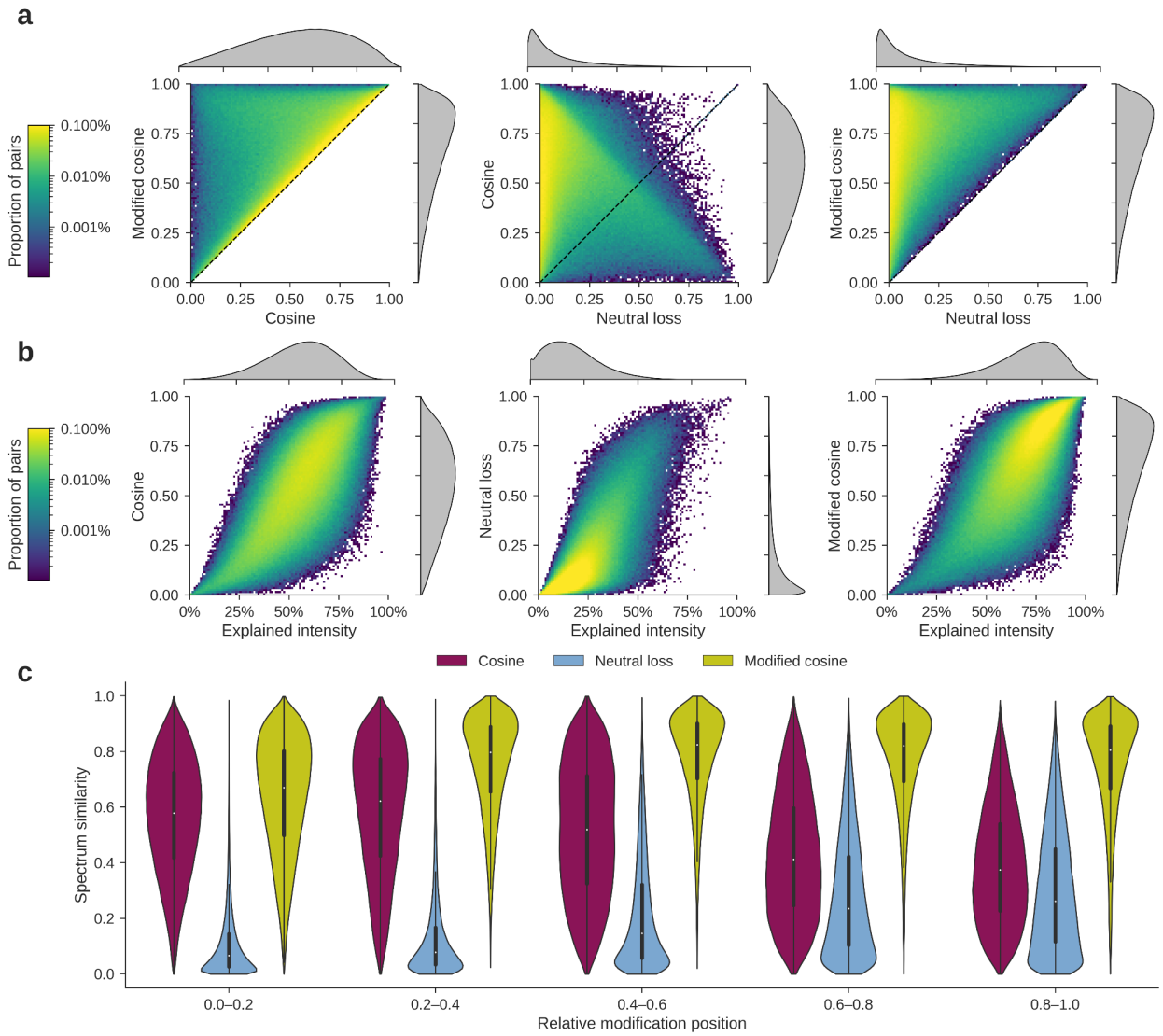

Figure 2. Comparison between the cosine similarity, neutral loss matching, and the modified cosine similarity when matching modified peptide MS/MS spectra. (a) Comparison between cosine similarity and modified cosine similarity (left), neutral loss matching and cosine similarity (middle), and neutral loss matching and modified cosine similarity (left). The heat maps show the relative number of spectrum pairs with the corresponding scores, out of all 955,228 peptide MS/MS spectrum pairs. The density plots (gray) show the distributions of the individual similarity measures. **(b)** Spectrum usage (explained intensity) for each of the spectrum similarity measures: Cosine similarity (left), neutral loss matching (middle), and modified cosine similarity (right). **(c)** Evaluation of the spectrum similarities in relation to the relative positions of the modifications. Modification positions are ordered N-terminal (0) to C-terminal (1).

Although pairs of peptide MS/MS spectra can serve as ground truth, because in this case all pairwise modifications are known, peptides are only a subset of molecules. Small molecules are chemically more diverse, and many of the MS/MS spectra will be significantly different. Therefore, we also evaluated the different spectrum similarity measures on a random subset of 10 millions spectrum pairs derived from the GNPS community spectral libraries (**Figure 3**).

Because these pairs were randomly selected, the majority of pairs exhibit low spectrum similarities as they correspond to molecules that are structurally unrelated. Similar to the previous analysis using peptide data, the modified cosine similarity strictly outperformed the cosine similarity and neutral loss matching (**Figure 3a**). Neutral loss matching performed considerably better than for peptide data, however, with 37.7% of the spectrum pairs achieving a higher neutral loss match score than cosine score, and 62.3% of the spectrum pairs achieving a higher cosine score than neutral loss score. Because this is a random set of mass spectrum comparisons where most are not expected to match, the spectrum usage is low for all three similarity measures (**Figure 3b**). Nevertheless, the rank of explained intensities, with the explained intensity from the modified cosine similarity outperforming the explained intensity from the cosine similarity, which in turn outperformed the explained intensity from neutral loss matching, was the same as for the peptide data. We also evaluated the spectrum similarity measures in function of the chemical similarity between the molecules, based on the Tanimoto index[24] as a proxy for structural similarity (**Figure 3c**). This indicated that although even for molecules with a high Tanimoto similarity the majority of pairs still exhibited a poor spectrum similarity, the modified cosine similarity was able to best reflect structural similarity.

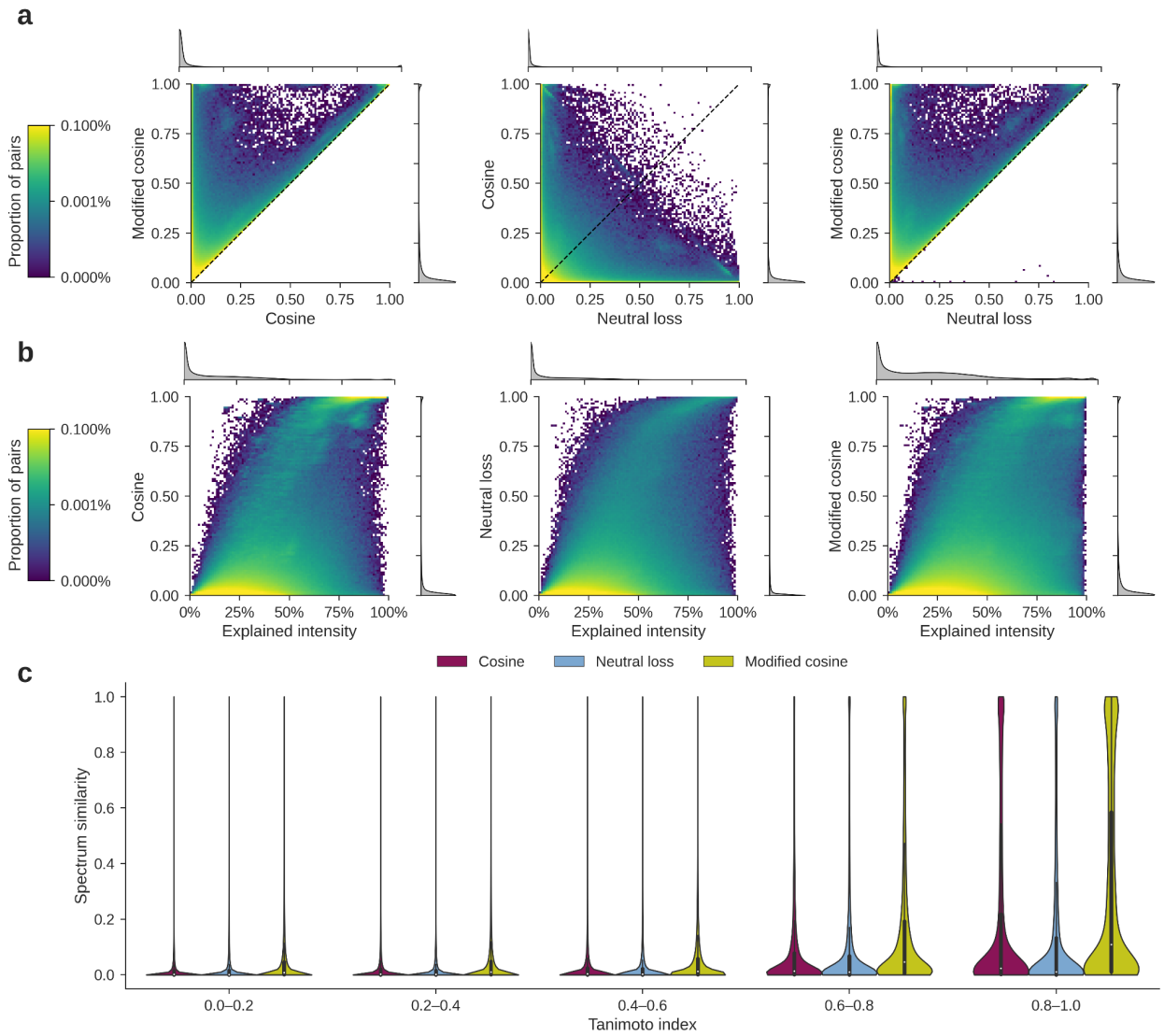

Figure 3. Comparison between cosine similarity, neutral loss matching, and modified cosine similarity when matching small molecule MS/MS spectra. (a) Comparison between cosine similarity and modified cosine similarity (left), neutral loss matching and cosine similarity (middle), and neutral loss matching and modified cosine similarity (left). The heat maps show the relative number of spectrum pairs with the corresponding scores, out of all 10 million small molecule MS/MS spectrum pairs. The density plots (gray) show the distributions of the individual similarity measures. **(b)** Spectrum usage (explained intensity) for each of the spectrum similarity measures: Cosine similarity (left), neutral loss matching (right), and modified cosine similarity (right). **(c)** Evaluation of the spectrum similarities in relation to the structural similarities, as measured using the Tanimoto index. A high Tanimoto index (0.8–1) indicates that the pair of molecules has high structural similarity.

Considering that the majority of the 10 million randomly selected spectrum pairs correspond to MS/MS spectra of unrelated molecules, we filtered for mass spectrum pairs whose molecules exhibit high structural similarity (Tanimoto index above 0.9; 6919 spectrum pairs). Although the

Tanimoto index is an imperfect method to fully assess structural similarity,[25] this filter still enriches for structurally related molecules. This data subset can provide additional insights into the impact of each spectrum similarity measure, especially when comparing the spectrum usage (**Figure 4**). This indicates that the modified cosine similarity captured more spectrum usage (as defined by explained intensity) than the cosine similarity and neutral loss matching. The different behavior of peptide spectrum pairs versus metabolite spectrum pairs, especially when comparing cosine similarity to neutral loss matching (**Figure 2a**, **Figure 3a**, **Figure 4**) suggests that various modifications may impact spectrum similarity differently.

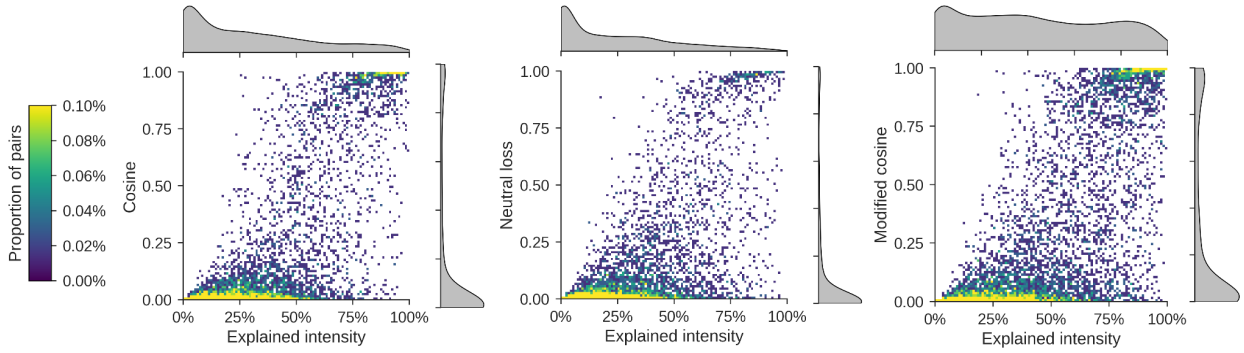

Figure 4. Comparison between cosine similarity, neutral loss matching, and modified cosine similarity when matching structurally similar (Tanimoto index above 0.9) small molecule MS/MS spectra. Spectrum usage (explained intensity) for each of the spectrum similarity measures: Cosine similarity (left), neutral loss matching (right), and modified cosine similarity (right).

To evaluate how specific types of modifications impact spectrum similarity, we performed detailed analyses of the bile acids molecular family, which has a number of distinct modifications. Our lab has recently discovered many new bile acids, for which the MS/MS spectra have been made available as an open access resource.[26–30] Additionally, we have collected MS/MS data on a historical library of previously determined bile acids. Although other studies have reported the recent discovery of new bile acids as well,[31–36] unfortunately these could not be included in the current analysis as the corresponding MS/MS data are not publicly available yet. Because our continued study of bile acids has given us a deep understanding of this subset of molecules, we can further analyze these data to understand how specific modifications influence spectrum similarity (**Figure 5**). In total 846 MS/MS spectra of 369 unique bile acids were included, leading to 340,637 pairs of MS/MS spectra. Although bile acids can undergo many different modifications,[37] here we focused on (i) all bile acid pairs, (ii) bile acid pairs that differ by a single oxygen, (iii) bile acid pairs that differ by a conjugated amino acid, and (iv) bile acids pairs that differ by the substitution of a conjugated amino acid. Similarly as for the previous analyses, the modified cosine similarity strictly outperformed neutral loss matching and the cosine similarity (**Figure 5a**). Interestingly, for this class of molecules neutral loss matching outperformed cosine similarity in 68.5% of cases.

An increasing Tanimoto index between bile acid pairs, reflecting more similarity in the structures, resulted in increased spectrum similarities as well, with the modified cosine similarity best

capturing high structural similarity (**Figure 5b**). However, these trends strongly depend on the type of modification. For a modification consisting of a single oxygen difference, the majority of spectrum pairs achieved high modified cosine similarities (median 0.807), whereas cosine similarities (median 0.430), and neutral loss scores (median 0.230) were considerably lower (**Figure 5c**). For comparing spectra of non-conjugated bile acids to conjugated bile acids, none of the spectrum similarity measures work well (**Figure 5d**). This indicates that this type of modification exerts a strong influence on the fragmentation pathways, despite belonging to the same class of molecules. Finally, when comparing spectra of conjugated bile acids that have undergone an amino acid substitution, neutral loss matching behaved most similarly to the modified cosine similarity, both showing a bimodal score distribution, and strongly outperforming the cosine similarity (**Figure 5e**). Nevertheless, for all of the mass spectrum pairs, there was not a single instance in which neutral loss matching outperformed the modified cosine similarity.

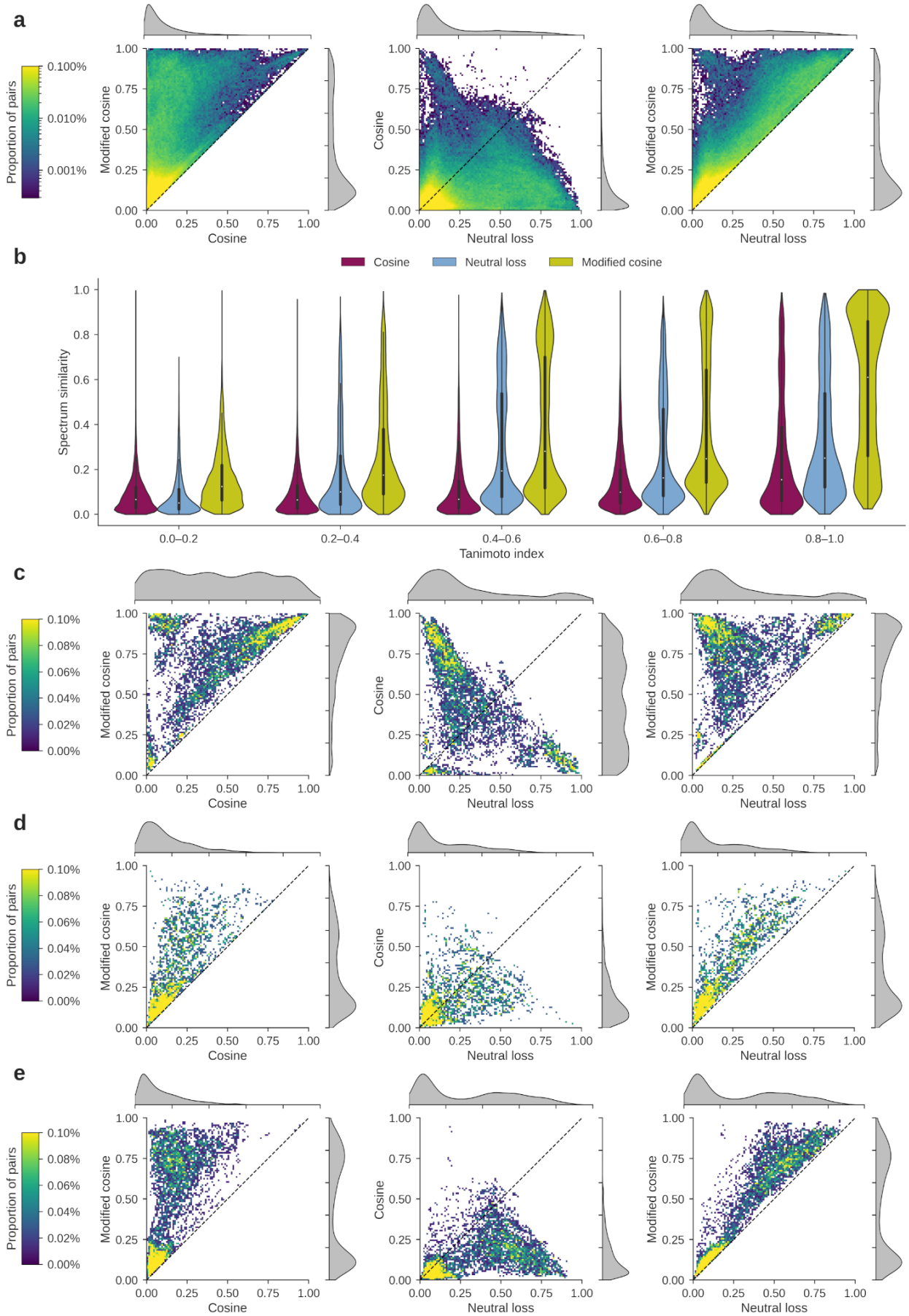
a
b
c
d
e
Proportion of pairs
0.100%
0.010%
0.001%
0.10%
0.08%
0.06%
0.04%
0.02%
0.00%
Modified cosine
Cosine
Neutral loss
Spectrum similarity
Tanimoto index
0.0–0.2
0.2–0.4
0.4–0.6
0.6–0.8
0.8–1.0

Figure 5. Comparison between cosine similarity, neutral loss matching, and modified cosine similarity when matching MS/MS spectra of bile acids. (a) Comparison between cosine similarity and modified cosine similarity (left), neutral loss matching and cosine similarity (middle), and neutral loss matching and modified cosine similarity (left). The heat maps show the relative number of spectrum pairs with the corresponding scores, out of all 340,637 bile acid MS/MS spectrum pairs. The density plots (gray) show the distributions of the individual similarity measures. **(b)** Evaluation of the spectrum similarities in relation to the structural similarities, as measured using the Tanimoto index. A high Tanimoto index (0.8–1) indicates that the pair of molecules has high structural similarity. **(c)** Comparison between spectrum similarity measures for 8128 bile acid pairs that differ by a single oxygen. **(d)** Comparison between spectrum similarity measures for 3515 bile acid pairs that differ by a conjugated amino acid. **(e)** Comparison between spectrum similarity measures for 8624 bile acid pairs that differ by the substitution of a conjugated amino acid.

Discussion

Here we have evaluated the performance of three related mass spectrum similarity measures—cosine similarity, neutral loss matching, and modified cosine similarity—in capturing similarities between structurally related molecules. Our evaluations indicate that the modified cosine similarity is superior to both alternative similarity measures for both peptide and small molecule MS/MS data. These results are in concordance with the popularity of the modified cosine similarity as one of the most commonly used methods to find MS/MS spectra of related molecules.[7]

Our interpretation from the peptide data is that the cosine similarity most closely approximates the modified cosine similarity when the modification is located in the N-terminal region of the peptide, as this will maximally conserve the overall spectral pattern of the dominant y-ions. Furthermore, despite only considering directly matching fragment ions, our results show that the traditional cosine similarity outperforms the recently proposed neutral loss matching strategy[20] in the majority of cases, even for modified small molecules. Despite the overall advantage of the cosine similarity, there are a significant number of small molecule MS/MS spectrum pairs for which neutral loss matching outperformed cosine similarity. In contrast, neutral loss matching was always inferior to the modified cosine similarity. We currently hypothesize that neutral loss matching outperforms the cosine similarity when the modification is on the same side of the molecule as the charge. In this case, all fragment ions are shifted by the modification mass, and the neutral loss spectra become mirror images of the original MS/MS spectra, whereas no fragment ions will match directly using the cosine similarity. Nevertheless, under such a circumstance the modified cosine similarity still captures all shifted fragment ions as well, which is confirmed by the comparisons presented. In conclusion, both cosine similarity and neutral loss matching can only capture a subset of fragment ions that are matched by the modified cosine similarity.

Although there is not a single test case where our analysis revealed neutral loss matching to outperform the modified cosine similarity, to enable community use of neutral loss matching as a comparative alternative to the cosine similarity and the modified cosine similarity, we have implemented it in several software tools, including matchms,[38] GNPS,[7] and MZmine.[39] For

detailed exploration of MS/MS spectrum similarity, user-friendly viewers are also available, for example within MZmine (**Figure 6**).[39,40] Additionally, to promote open and reproducible science, all data and code to execute the presented analyses are available with an open and unrestricted license.

Figure 6. Viewer in MZmine for user-friendly evaluation of spectrum similarity measures. The example shows the spectrum usage from the modified cosine similarity (top) and neutral loss matching (bottom) for the MS/MS spectra of the bile acids taurocholic acid (CCMSLIB00005435561) and glycocholic acid (CCMSLIB00005435513). The viewer shows the directly matching fragment ions (top) and neutral loss ions (bottom) in blue, and the matching fragment ions that are offset by the precursor mass difference (top) in orange. The side panel shows the relative contribution of all pairs of matching fragment and neutral loss ions to the overall score.

Besides the three mass spectrum similarity measures that we have evaluated in this work, an increasing number of other spectrum similarity measures for the discovery of MS/MS spectra of related molecules are being proposed, including various MS/MS spectrum preprocessing steps, different implementations of (modified) cosine similarities that utilize different weighting schemes,[11,19,41] and methods that use machine learning, such as Spec2Vec,[14] MS2DeepScore,[16] spectral entropy,[15] SIMILE,[17] GLEAMS,[18] and many others. These similarity measures likely provide alternative and complementary approaches for the discovery of related molecules, as the modified cosine similarity only accounts for mass shifts associated with a single structural modification, whereas in practice the structural relationship between two analogs can consist of two or more modifications, resulting in more complex mass fragmentation relationships. Future

benchmarking of such methods, as well as the effects of data processing, are necessary to provide further insight into the strengths and limitations of each approach.

Acknowledgments

This research was supported by BBSRC-NSF award 2152526 and National Institutes of Health award U19 AG063744.

Author contributions statement

PCD conceptualized and supervised the work. WB, RS, and FH implemented various spectrum similarity measures. WB, RS, JJJvdH, MW, and PCD performed the data analyses. MW provided computational resources. WB and PCD wrote the manuscript. All authors reviewed and edited the manuscript.

Competing interests statement

PCD is on the advisory board of Cybele and is a co-founder and advisor of Ometa and Enveda, with prior approval by UC San Diego. JJJvdH is a member of the Scientific Advisory Board of NAICONS Srl., Milan, Italy. MW is a co-founder of Ometa Labs LLC.

Methods

Spectrum similarity measures

The cosine similarity, modified cosine similarity, and neutral loss matching were implemented in Python (version 3.10). In contrast to the original formulation of neutral loss spectra,[20] spectra were not mirrored at their precursor *m*/*z*, as this is not properly defined for multiply charged precursors (i.e. a multiply charged precursor peak can have a lower *m*/*z* value than its singly charged fragment ions), but instead at the theoretical *m*/*z* of the singly charged precursor. For the cosine similarity and neutral loss matching, only directly matching fragment ions with a near-identical *m*/*z* or $\Delta m/z$, respectively, were considered. For the modified cosine similarity, fragment ions with *m*/*z* values that are shifted by the pairwise precursor mass difference were considered as well. For all similarity measures, the optimal matching peak assignment across both MS/MS spectra was computed using the SciPy (version 1.8.0)[42] implementation of the linear assignment problem. Additional scientific computing used Numba (version 0.55.1),[43] NumPy (version 1.21.5),[44] and spectrum_utils (version 0.4.0)[45] for efficient spectrum similarity computations, Matplotlib (version 3.5.1)[46] and Seaborn (version 0.11.2)[47] for data visualization, Pyteomics (version 4.5.3)[48] for data loading, and Jupyter notebooks[49] and Pandas (version 1.4.1)[50] for data analysis.

Analysis of modified peptide MS/MS spectrum pairs

Pairs of modified peptide MS/MS spectra were extracted from the MassIVE-KB spectral library (version 2018/06/15).[21] This is a repository-wide spectral library derived from over 30 TB of

human MS/MS proteomics data, containing 2,154,269 MS/MS spectra. First, MS/MS spectra with precursor charge 2, 3, or 4 were considered and spectrum pairs were required to have identical precursor charges. Second, pairs of spectra whose peptides differ by a single modification were extracted. This modification can consist of the absence/presence of a single post-translational modification, or a single amino acid difference (edit distance of 1, corresponding to a single amino acid substitution or amino acid prefix or suffix addition/removal). To avoid inclusion of MS/MS spectrum pairs that only differ by nearby ambiguously localized post-translational modifications, all MS/MS spectrum pairs were required to have a precursor mass difference of at least 4 *m*/*z*. This resulted in 955,228 MS/MS spectrum pairs with known peptide labels derived from the MassIVE-KB spectral library.

MS/MS spectra were preprocessed by removing peaks within a 0.1 *m*/*z* window around the precursor *m*/*z*. Spectrum similarities were computed using a 0.1 *m*/*z* fragment mass tolerance. The relative locations of the modifications by which pairs of peptides differ were calculated by normalizing the modification indexes in each pair of peptide sequences by the length of the shortest paired peptide sequence.

Analysis of modified small molecule MS/MS spectrum pairs

Pairs of modified small molecule MS/MS spectra were extracted from the GNPS community spectral libraries (ALL_GNPS_NO_PROPOGATED), consisting of 495,600 MS/MS spectra, and the GNPS bile acids library (BILELIB19), consisting of 4533 MS/MS spectra (data downloaded on May 12, 2022).[7] First, MS/MS spectra with precursor charge 1 were considered (spectra with the unspecified precursor charge 0 were assumed to have precursor charge 1 as well). Additionally, only centroid MS/MS spectra that contain at least 6 fragment ions, were acquired in positive ion mode, and have [M+H]+ adducts were included.

For the general small molecule analysis, all MS/MS spectrum pairs from the GNPS community spectral libraries with a pairwise precursor mass difference between 1 and 200 *m*/*z* were considered, resulting in over 1.5 billion possible spectrum pairs. In deference to computational limitations, 10 million MS/MS spectrum pairs were randomly selected for the subsequent analysis.

For the modified bile acids analysis, all MS/MS spectrum pairs from the GNPS bile acids library with a pairwise precursor mass difference between 1 and 200 *m*/*z* were considered, resulting in 340,637 spectrum pairs. Additionally, bile acids that differ by a single oxygen were selected by filtering on a pairwise precursor mass difference (mass tolerance 0.01 *m*/*z*) of 15.9949 *m*/*z*; bile acids that differ by a conjugation were selected by filtering on pairwise precursor mass differences (mass tolerance 0.01 *m*/*z*) of 57.0214 *m*/*z* (glycine), 71.0371 *m*/*z* (alanine), 107.0041 *m*/*z* (taurine), 147.0684 *m*/*z* (phenylalanine), or 163.0633 *m*/*z* (tyrosine); and bile acids that differ by a substitution in conjugation were selected by filtering on pairwise precursor mass differences (mass tolerance 0.01 *m*/*z*) of 49.9826 *m*/*z* (taurine ↔ glycine), 35.9669 *m*/*z* (tauro ↔ alanine), 40.0643 *m*/*z* (taurine ↔ phenylalanine), 56.0592 *m*/*z* (taurine ↔ tyrosine),

90.0469 *m/z* (glycine ↔ phenylalanine), 106.0418 *m/z* (glycine ↔ tyrosine), 76.0313 *m/z* (alanine ↔ phenylalanine), or 92.0262 *m/z* (alanine ↔ tyrosine).

MS/MS spectra were preprocessed by removing peaks within a 0.1 *m/z* window around the precursor *m/z*. Spectrum similarities were computed using a 0.1 *m/z* fragment mass tolerance. Tanimoto indexes[24] were computed using RDKit (version 2022.3.2)[51] by creating molecule objects from their SMILES strings, deriving RDKit topological fingerprints using the RDKit default settings (2048 bits), and computing the Tanimoto index.

Data and code availability

All code and notebooks to recreate the presented analyses are available as open source under the permissive BSD license at https://github.com/bittremieux/cosine_neutral_loss. A permanent archive of the source code and the analysis notebooks is available on Zenodo at https://doi.org/10.5281/zenodo.6584619. Additionally, cosine similarity, modified cosine similarity, and neutral loss matching have been implemented in the matchms Python package (https://github.com/matchms/matchms; version 0.15.0).[38] An interactive viewer to inspect neutral loss matching and (modified) cosine similarity is available in MZmine (https://github.com/mzmine/mzmine3)[39] under "Menu > Tools > Spectral mirror." Spectra can be retrieved by their GNPS library identifiers or Universal Spectrum Identifiers (USIs).[40,52] Experimental MS/MS spectra can be selected from feature finding results by selecting two rows in a feature table and choosing "Show > MS/MS spectral mirror" from the right-click context menu.

The MassIVE-KB spectral library and the GNPS community spectral libraries are available under an open and free license at https://massive.ucsd.edu/ProteoSAFe/static/massive-kb-libraries.jsp and https://gnps-external.ucsd.edu/gnpslibrary, respectively. A permanent archive of the spectral libraries, as well as the computed similarity scores for the different analyses, is available on Zenodo at https://doi.org/10.5281/zenodo.6584635.

Individual spectra are accessible by their Universal Spectrum Identifiers (USIs).[40,52] The spectra displayed in **Figure 1** and **Figure 6** are:

- Apratoxin A: mzspec:GNPS:GNPS-LIBRARY:accession:CCMSLIB00000424840
- Apratoxin A analog with loss of 30.010 Da: mzspec:MSV000086109:BD5_dil2x_BD5_01_57213:scan:760
- Taurocholic acid: mzspec:GNPS:GNPS-LIBRARY:accession:CCMSLIB00005435561
- Glycocholic acid: mzspec:GNPS:GNPS-LIBRARY:accession:CCMSLIB00005435513

References

1. Watrous, J. *et al.* Mass spectral molecular networking of living microbial colonies. *Proc. Natl. Acad. Sci.* **109**, E1743–E1752 (2012).
2. Stein, S. Mass spectral reference libraries: An ever-expanding resource for chemical identification. *Anal. Chem.* **84**, 7274–7282 (2012).
3. Vinaixa, M. *et al.* Mass spectral databases for LC/MS- and GC/MS-based metabolomics: State of the field and future prospects. *TrAC Trends Anal. Chem.* **78**, 23–35 (2016).
4. Beniddir, M. A. *et al.* Advances in decomposing complex metabolite mixtures using substructure- and network-based computational metabolomics approaches. *Nat. Prod. Rep.* **38**, 1967–1993 (2021).
5. Stein, S. E. & Scott, D. R. Optimization and testing of mass spectral library search algorithms for compound identification. *J. Am. Soc. Mass Spectrom.* **5**, 859–866 (1994).
6. Wan, K. X., Vidavsky, I. & Gross, M. L. Comparing similar spectra: From similarity index to spectral contrast angle. *J. Am. Soc. Mass Spectrom.* **13**, 85–88 (2002).
7. Wang, M. *et al.* Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nat. Biotechnol.* **34**, 828–837 (2016).
8. Moorthy, A. S., Wallace, W. E., Kearsley, A. J., Tchekhovskoi, D. V. & Stein, S. E. Combining fragment-ion and neutral-loss matching during mass spectral library searching: A new general purpose algorithm applicable to illicit drug identification. *Anal. Chem.* **89**, 13261–13268 (2017).
9. Aron, A. T. *et al.* Reproducible molecular networking of untargeted mass spectrometry data using GNPS. *Nat. Protoc.* **15**, 1954–1991 (2020).
10. Bandeira, N., Tsur, D., Frank, A. & Pevzner, P. A. Protein identification by spectral networks analysis. *Proc. Natl. Acad. Sci.* **104**, 6140–6145 (2007).
11. Burke, M. C. *et al.* The hybrid search: A mass spectral library search method for discovery of modifications in proteomics. *J. Proteome Res.* (2017) doi:10.1021/acs.jproteome.6b00988.
12. Bittremieux, W., Meysman, P., Noble, W. S. & Laukens, K. Fast open modification spectral library searching through approximate nearest neighbor indexing. *J. Proteome Res.* **17**, 3463–3474 (2018).
13. Bittremieux, W., Laukens, K. & Noble, W. S. Extremely fast and accurate open modification spectral library searching of high-resolution mass spectra using feature hashing and graphics processing units. *J. Proteome Res.* **18**, 3792–3799 (2019).
14. Huber, F. *et al.* Spec2Vec: Improved mass spectral similarity scoring through learning of structural relationships. *PLOS Comput. Biol.* **17**, e1008724 (2021).
15. Li, Y. *et al.* Spectral entropy outperforms MS/MS dot product similarity for small-molecule compound identification. *Nat. Methods* **18**, 1524–1531 (2021).
16. Huber, F., van der Burg, S., van der Hooft, J. J. J. & Ridder, L. MS2DeepScore: A novel deep learning similarity measure to compare tandem mass spectra. *J. Cheminformatics* **13**, 84 (2021).
17. Treen, D. G. C. *et al.* SIMILE enables alignment of tandem mass spectra with statistical significance. *Nat. Commun.* **13**, 2510 (2022).
18. Bittremieux, W., May, D. H., Bilmes, J. & Noble, W. S. A learned embedding for efficient joint

analysis of millions of mass spectra. *Nat. Methods*, (2022).
19. Xue, J., Guijas, C., Benton, H. P., Warth, B. & Siuzdak, G. METLIN MS2 molecular standards database: A broad chemical and biological resource. *Nat. Methods* **17**, 953–954 (2020).
20. Aisporna, A. *et al.* Neutral loss mass spectral data enhances molecular similarity analysis in METLIN. *J. Am. Soc. Mass Spectrom.* **33**, 530–534 (2022).
21. Wang, M. *et al.* Assembling the community-scale discoverable human proteome. *Cell Syst.* **7**, 412-421.e5 (2018).
22. Renard, B. Y. *et al.* When less can yield more - Computational preprocessing of MS/MS spectra for peptide identification. *PROTEOMICS* **9**, 4978–4984 (2009).
23. Bittremieux, W. *et al.* Open access repository-scale propagated nearest neighbor suspect spectral library for untargeted metabolomics. *bioRxiv* (2022) doi:10.1101/2022.05.15.490691.
24. Tanimoto, T. T. *An elementary mathematical theory of classification and prediction*. https://scholar.google.com/scholar_lookup?title=elementary+mathematical+theory+of+classification+and+prediction&author=Tanimoto%2C+T.+T.&publication_year=1958 (1958).
25. Bajusz, D., Rácz, A. & Héberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J. Cheminformatics* **7**, 20 (2015).
26. Quinn, R. A. *et al.* Global chemical effects of the microbiome include new bile-acid conjugations. *Nature* **579**, 123–129 (2020).
27. Dethloff, F. *et al.* Paroxetine administration affects microbiota and bile acid levels in mice. *Front. Psychiatry* **11**, 518 (2020).
28. Petras, D. *et al.* Chemical proportionality within molecular networks. *Anal. Chem.* **93**, 12833–12839 (2021).
29. Hoffmann, M. A. *et al.* High-confidence structural annotation of metabolites absent from spectral libraries. *Nat. Biotechnol.* **40**, 411–421 (2021).
30. Gentry, E. *et al.* A synthesis-based reverse metabolomics approach for the discovery of chemical structures from humans and animals. *Res. Sq.* (2021) doi:10.21203/rs.3.rs-820302/v1.
31. Lucas, L. N. *et al.* Dominant bacterial phyla from the human gut show widespread ability to transform and conjugate bile acids. *mSystems* **6**, e00805-21 (2021).
32. Zhu, Q.-F. *et al.* Alternating dual-collision energy scanning mass spectrometry approach: Discovery of novel microbial bile-acid conjugates. *Anal. Chem.* **94**, 2655–2664 (2022).
33. Garcia, C. J., Kosek, V., Beltrán, D., Tomás-Barberán, F. A. & Hajslova, J. Production of new microbially conjugated bile acids by human gut microbiota. *Biomolecules* **12**, 687 (2022).
34. Ma, Y. *et al.* BAFinder: A software for unknown bile acid identification using accurate mass LC-MS/MS in positive and negative modes. *Anal. Chem.* **94**, 6242–6250 (2022).
35. Shalon, D. *et al.* Profiling of the human intestinal microbiome and bile acids under physiologic conditions using an ingestible sampling device. *bioRxiv* (2022) doi:10.1101/2022.01.19.476920.
36. Neugebauer, K. A. *et al.* Bile acid-CoA:amino acid N-acyltransferase gene knockout alters early life development, the gut microbiome and reveals unusual bile acid conjugates in mice. *bioRxiv* (2022) doi:10.1101/2022.04.10.487642.
37. Hofmann, A. F. & Hagey, L. R. Key discoveries in bile acid chemistry and biology and their

clinical applications: history of the last eight decades. *J. Lipid Res.* **55**, 1553–1595 (2014).
38. Huber, F. *et al.* matchms - processing and similarity evaluation of mass spectrometry data. *J. Open Source Softw.* **5**, 2411 (2020).
39. Pluskal, T., Castillo, S., Villar-Briones, A. & Orešič, M. MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics* **11**, 395 (2010).
40. Bittremieux, W. *et al.* Universal MS/MS visualization and retrieval with the Metabolomics Spectrum Resolver web service. *bioRxiv* (2020) doi:10.1101/2020.05.09.086066.
41. Chen, L. *et al.* Metabolite discovery through global annotation of untargeted metabolomics data. *Nat. Methods* **18**, 1377–1385 (2021).
42. SciPy 1.0 Contributors *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* (2020) doi:10.1038/s41592-019-0686-2.
43. Lam, S. K., Pitrou, A. & Seibert, S. Numba: a LLVM-based Python JIT compiler. in *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC - LLVM '15* 1–6 (ACM Press, 2015). doi:10.1145/2833157.2833162.
44. Harris, C. R. *et al.* Array programming with NumPy. *Nature* **585**, 357–362 (2020).
45. Bittremieux, W. spectrum_utils: A Python package for mass spectrometry data processing and visualization. *Anal. Chem.* **92**, 659–661 (2020).
46. Hunter, J. D. Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).
47. Waskom, M. L. seaborn: Statistical data visualization. *J. Open Source Softw.* **6**, 3021 (2021).
48. Levitsky, L. I., Klein, J. A., Ivanov, M. V. & Gorshkov, M. Pyteomics 4.0: Five years of development of a Python proteomics framework. *J. Proteome Res.* **18**, 709–714 (2019).
49. Thomas, K. *et al.* Jupyter Notebooks -- A publishing format for reproducible computational workflows. in *Positioning and Power in Academic Publishing: Players, Agents and Agendas* 87–90 (IOS Press, 2016).
50. McKinney, W. Data structures for statistical computing in Python. in *Proceedings of the 9th Python in Science Conference* (eds. van der Walt, S. & Millman, J.) 51–56 (2010).
51. Landrum, G. *et al. rdkit/rdkit: 2022_03_2 (Q1 2022) Release*. (Zenodo, 2022). doi:10.5281/ZENODO.6483170.
52. Deutsch, E. W. *et al.* Universal Spectrum Identifier for mass spectra. *Nat. Methods* **18**, 768–770 (2021).