Tailored mass spectral data exploration using the specXplore interactive dashboard - Supplementary Information

Kevin Mildau[1,2,3,*], Henry Ehlers[4,*], Ian Oesterle[2,5,6], Manuel Pristner[2,5], Benedikt Warth[5], Maria Doppler[7], Christoph Bueschl[7], Juergen Zanghellini[1,**], and Justin J.J van der Hooft[8,9,**]

[1]Biochemical Network Analysis Lab, Department of Analytical Chemistry, University of Vienna, Vienna, Austria
[2]Doctoral School in Chemistry (DOSCHEM), University of Vienna, 1090 Vienna, Austria
[3]Austrian Centre of Industrial Biotechnology (ACIB GmbH), 8010 Graz, Austria
[4]Institute of Visual Computing and Human-Centered Technology, TU Wien, 1040 Vienna, Austria
[5]Department of Food Chemistry and Toxicology, University of Vienna, 1090 Vienna, Austria.
[6]Department of Biophysical Chemistry, University of Vienna, 1090 Vienna, Austria
[7]University of Natural Resources and Life Sciences (BOKU), 3430 Tulln, Austria
[8]Bioinformatics Group, Wageningen University, 6708PB Wageningen, The Netherlands
[9]Department of Biochemistry, University of Johannesburg, 2006 Johannesburg, South Africa
[*]Shared First Authors.
[**]Corresponding Authors: juergen.zanghellini@univie.ac.at, justin.vanderhooft@wur.nl

March 20, 2024

Contents

1 Technical Terms

In the main manuscript a number of technical terms are used. For the interested but unfamiliar reader, we herewith include a number of term clarifications and pointers to clarify our choice of language.

- A network is a collection of connected features. In our case, a network consists of MS/MS spectral features connected provided their spectral similarity is high. Networks are represented using node-link-diagrams.
- Node-link diagram - a term commonly used to refer to the graphical representation of a network via nodes and links (i.e. edges). In this paper, we use node-link diagram and network-view interchangeably.
- A node is a feature in a network that can be connected to other features via edges. An alternative term for node is vertex.
- An edge is a connection between two nodes. Other terms for edges are links or vertices.
- Network layout refers to the spatial arrangement of nodes and edges on an usually two dimensional plotting surface. Network layout is also sometimes referred to as embedding. This term is avoided in this paper to avoid confusion with embedding in the machine learning sense.
- Given a network $G(V, E)$, where V denotes its nodes and E its (weighted) edges, we define its topology as the relationships between individual (groups of) nodes and edges or the network as a whole, irrespective of the network's layout.
- Molecular Networking (MN) is an exploratory data analysis technique merging spectral similarity-based topological clustering and visualization as node-link diagrams.
- The plain English words group/grouping are wherever appropriate to avoid jargon terms such as clustering (as in k-medoid or k-means clustering), embedding (as in projection of groups of features into a close-by lower dimensional space), or molecular families. The latter are groups of spectral data features clustered and visualized as network-views via traditional MN or feature based molecular networking (FBMN). Molecular families, usually represent smaller, disconnected networks that are part of a larger dataset. When we refer to this disconnected nature, we use the phrasing disjoint sub-network for emphasis.

2 Importing

2.1 Spectral data pre-processing

SpecXplore makes use of pairwise similarity computations and direct spectral comparison approaches that require some spectral data pre-processing. First, spectra are normalized such that intensities are measured relative to the most intense fragment. This is necessary for some pairwise similarity matrices, and fragmap overviews. Subsequently, the spectra are filtered to only keep fragments if their relative intensity exceeds some set minimum or to limit the spectrum to the top n (default of two-hundred) fragments ranked by fragment intensity. This step is done in order to a) limit noise peaks from diluting signals that interfere with pairwise spectral similarity measure computations, and b) limit very large and noisy spectra from diluting and obscuring fragmentation overview visualizations while contributing little effective insight. Finally, any fragments outside the mass-to-charge ratio window of zero to one thousand will be removed as they are not supported by the used machine learning model and fall outside the spectral binning range. Additional processing such a minimum number of fragments per spectrum may also be applied to focus analysis on more promising and information rich spectra for analysis. It should be noted that these processing steps happen in specXplore on the basis of .mgf spectral exports. Whenever possible, pre-processing should be done in software operating closer to the raw data, such as e.g. MZMine [1], as more relevant information for filtering and quality checks will be available.

2.2 Usage and tuning of t-SNE as a fixed network layout

Finding good settings for t-SNE is a task that is difficult to automate and may even warrant a dashboard of its own to formally evaluate and understand as done in t-viSNE [2]. In our trials, we've noticed that often times very high perplexities lead to the best distance preservation for the t-SNE embedding. However, larger perplexity usually implies larger groups of nodes and hence larger overlaps between topological groups. Hence, the best perplexity settings for t-SNE require an informal trade-off in distance preservation and visual characteristics. Despite lacking formally optimal settings for the used t-SNE embedding, we explicitly chose to make the t-SNE embedding in specXplore fixed rather than modifiable on the fly. This has three reasons; computational speed, robustness, and mental map preservation. First, t-SNE embedding computations can be computationally costly for larger datasets and may require long run-times. In specXplore and dash, long run times may be easy to confuse with app crashes. In addition, t-SNE embeddings may require a number of t-SNE runs to be able to assess what embedding works well in a qualitative sense. This effectively multiplies the computational speed bottleneck of t-SNE tuning in the app. These two aspects together make tuning prohibitive in an interactive session. Second, during our trials t-SNE runs showed surprising qualitative robustness, often leading to useful embeddings regardless of precise perplexity values used. Third, the t-SNE embedding is meant to provide a global overview from which to explore the dataset. Should nodes be scattered differently, the user's 'mental map' and hence overview of the data would be lost. Given these considerations, it seemed most appropriate to defer t-SNE tuning to the jupyter notebook pre-processing step. Alternative approaches to data visualization were considered such as tmap [3, 4] or pacmap [5]. However, given the precedent of t-SNE in the field for instance MetGem [6], as well as a combination of good agreement with local topology at higher thresholds and straightforward compatibility with arbitrary similarity matrices, made t-SNE our method of choice. Alternative embedding approaches, include those approaches that avoid the computation of a full pairwise similarity matrix, may be especially useful in larger visualization contexts going beyond individual experiment visualization.

2.3 In-silico spike-in standards

SpecXplore includes the option of using in-silico spike-in reference standards in the spectral data exploration workflow. This is done via adding additional spectra and corresponding metadata and assigning the add-on features as features to be-highlighted in the t-SNE overview. This option combines local library search and molecular networking style visualization in a single exploratory visualization approach. Naturally, such standards can be from any online repository and instrument type, but can also be from in-house libraries affording greater spectral similarity through similar or equal instrumental setups being used in standard and experimental spectra acquisition. By making reference standards stand out visually using higher opacity and darker grays, chemical spaces of interest can be quickly identified in the global overview.

2.4 The default similarity scores in specXplore

The primary similarity score used in data exploration is ms2deepscore, which computes pairwise spectral similarities by transforming any two input spectra into the models' deep-learning-based embedding representation [7]. This step involves binning of spectra and recasting of binned spectra in the embedding space. Embedded spectra are compared against each other using dot product similarities. Importantly, the model is trained to generate embeddings such that the resulting dot products of two generated embeddings resemble structural tanimoto score based similarities as closely as possible. SpecXplore makes use of the pre-trained ms2deepscore model also used in ms2query, circumventing the need for users to train and validate their own models [8]. In addition to ms2deepscore pairwise similarities, modified cosine score similarities and pre-trained spec2vec embedding based similarities, are computed and integrated within specXplore for comparative purposes [9, 10]. Especially the comparison between ms2deepscore and modified cosine scores can be insightful to assess whether the machine learning model provides additional, or fails to detect fragmentation based associations. Alternative scores could be used in specXplore, but would have to be provided by the user and care needs to be taken that they lie in the range between 0 and 1 for the default distance matrix conversion steps in specXplore to be applicable.

2.5 Chemical Class Prediction Integration

An optional but recommended part of specXplore is to run ms2query to provide analog library matches for all experimental spectra. ms2query makes use of a local library against which experimental spectra are compared using optimized machine learning-based matching. ms2query putative structure annotations serve as additional metadata available for each spectrum within the specXplore dashboard. Putative structures, but crucially also the chemical ontology corresponding to the determined analog can give indications about the nature of the compound represented by the spectrum in question. Taking this approach further, the analog classifications may be supplied to specXplore's classification table, providing a rough, indirect analog based classification for all experimental spectra that can be used to identify chemical regions of interest. Of course, other tools for library matching or more direct class predictions using tools such as CANOPUS may also be used and integrated into the metadata table and classification tables of specXplore [11, 12]. Integration of these tools' class predictions can be achieved easily using specXplore's provided a) the chemical classification is available in tabular form and b) feature identifiers matching those supplied with the mgf file are available for each classification.

3 Results

3.1 Wheat Data Example

This section contains additional details regarding the wheat data [13, 14] as well as additonal figures.

3.1.1 Raw MS/MS Data Processing

All raw files were processed with MZmine 3 (version 3.3.0). Data was loaded in both ionization modes separately [1]. Then mass detection was done with a noise level of 1E4, followed by the ADAP chromatogram builder module (min. number of scans: 4, min. highest intensity: 1E5, scan to scan accuracy: 5 ppm). The chromatograms were smoothed with a Savitzky Golay filter (width 7 points), and chromatographic peaks were detected with the Local minimum feature resolver module (Chromatographic threshold: 10%, minimum absolute height: 1E5, Min. ratio of top to edge: 5, Peak duration: 0.05 − 2 min). The sample specific feature lists were then filtered for only monoisotopic features (Isotopic peaks finder module), and all feature lists were combined using the join aligner module (m/z tolerance: 10 ppm, Retention time tolerance: 0.2 min). Then the combined feature list was filtered for chromatographic peaks with at least one assigned MSMS scans only and all other features were discarded (module Feature list rows filter). Next, the data was exported to the MGF file format using the GNPS export module before further processing with specXplore. Reference standard spectra [15] where spiked into the dataset via inclusion in the MGF file.

3.1.2 Quantitative Similarity via Augmap

Visualization of the pairwise similarity matrix and alternative scores using AUGMAP provides a means of quantitative evaluation of spectral similarities for subselections of the data (Figure 1)

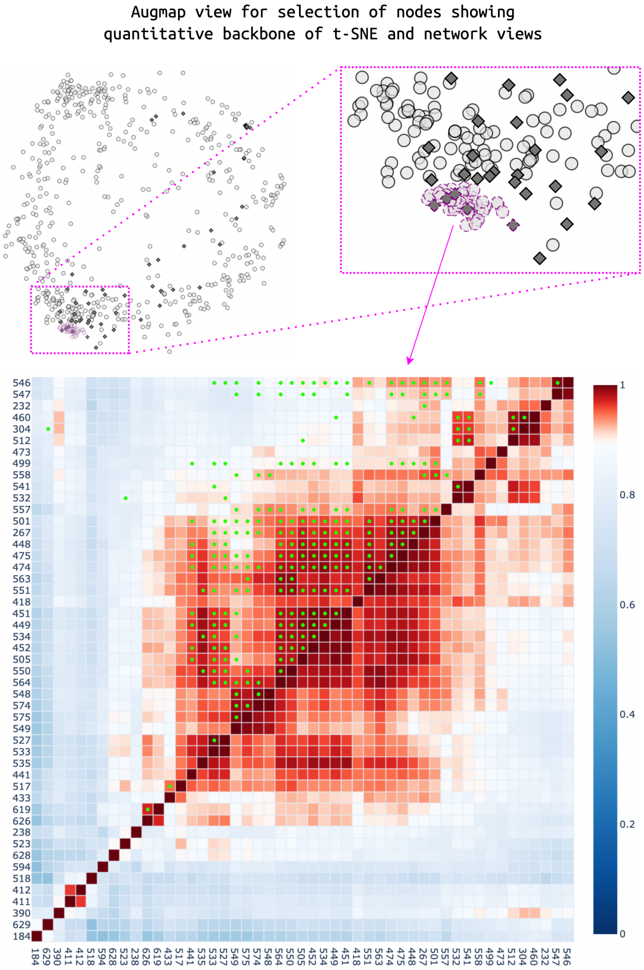
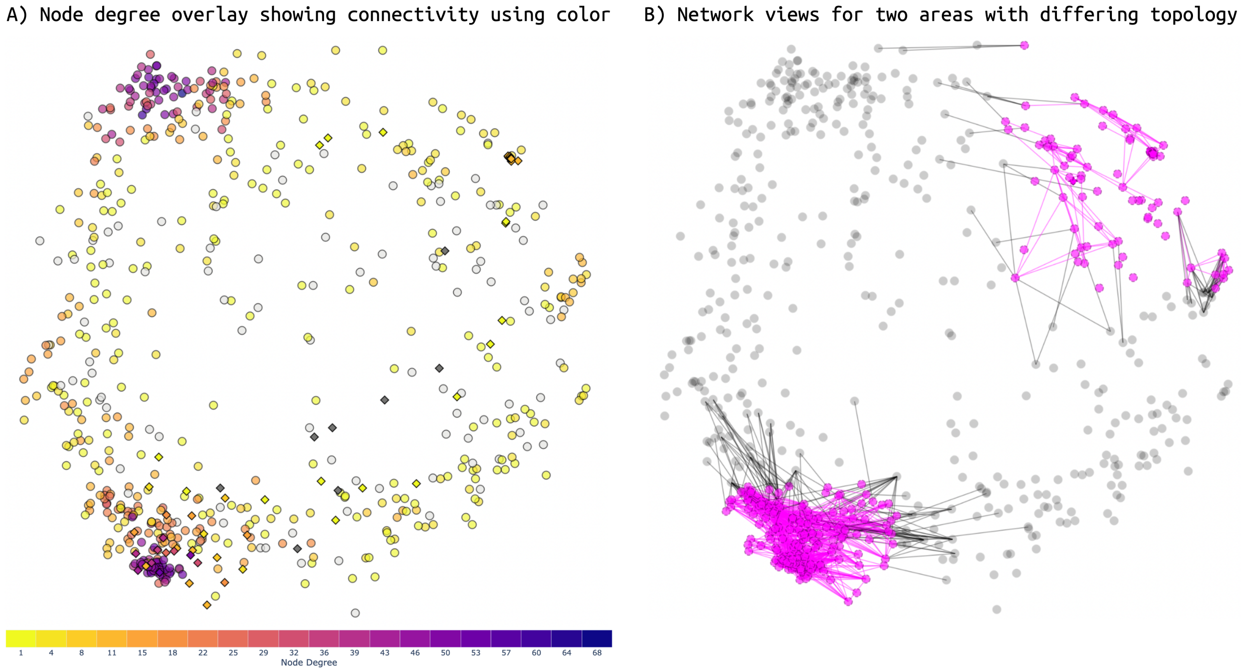


Figure 1: Pairwise spectral similarity and score adjacency matrix agreement visualized as marker augmented heatmaps. Feature identifiers are repeated on x and y axis, while the color gradient indicates quantitative similarity for the primary score, i.e. ms2deepscore. Here, a small selection of nodes in the bottom left area was selected. At threshold level of 0.9 for ms2deepscores, many of the selected nodes are connected via edges, indicated as red in the divergent color scale around the threshold. The remainder of the nodes tend towards light blue and white color, indicating that lowering threshold to 0.8 or 0.7 would lead to almost all pairwise connections being above threshold. In addition, we can see that at current numeric threshold value of 0.9, adjacency matrix (i.e. which pairwise relationships form edges or not) disagreement is high via the lack of or presence of circular (modified cosine score) and rectangular (spec2vec) markers. In this area of the graph, ms2deepscore appears to find the strongest connectivity patterns, while its adjacency matrix is missing some of the cosine score connections at threshold 0.9. No spec2vec connections are present at current thresholds, indicating that the pre-trained model covers the selected node subset poorly or that the score requires lower thresholds for connectivity to become apparent. Hover pop-ups are available to assess exact scores across the three scoring approaches for quantitative insight.

3.1.3 Node Topology Views

SpecXplore provides two general ways of assessing topological information via node degree colors overlays and network overlays, here illustrated with the wheat data in Figure 2.



A) Node degree overlay showing connectivity using color

B) Network views for two areas with differing topology

Figure 2: A) Node degree visualization for threshold settings of 0.8. Color indicates node degree in qualitative brackets of 10, going from lowest node degree to the highest. At this threshold, the highest node degree is 68 while the lowest visually highlighted degree is 1. Node degrees of zero are off color scale via grey color. Regions with high network connectivity are clearly visible, as are changes in topology in response to threshold changes. B) At identical threshold levels different parts of the data exhibit vastly different connectivity. For example, at threshold levels of 0.8 the lower left corner of the two-dimensional embedding is densely interconnected, while the top right corner shows sparse connectivity.

3.1.4 Classification-Based Overviews

SpecXplore provides two means of highlighting groups of spectra of potential interest in the form of A) k-medoid cluster coloring, and B) chemical class based color highlighting. The two are illustrated in Figure 3 for the wheat data.

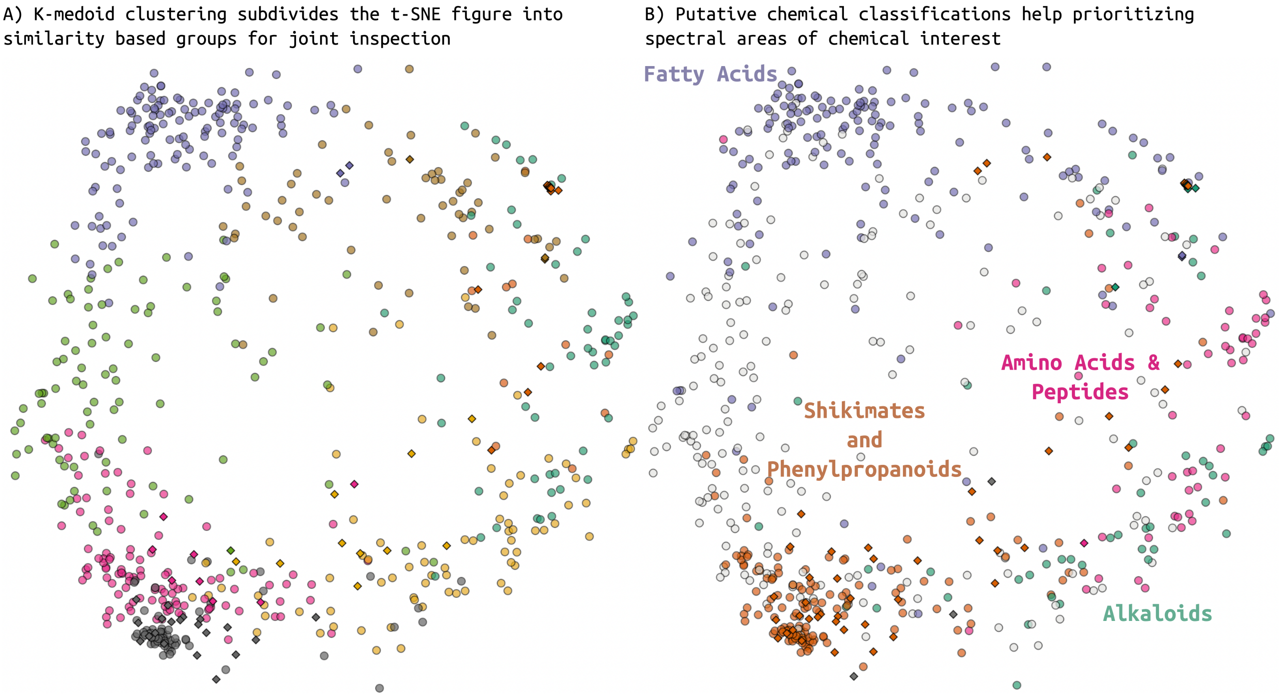


Figure 3: SpecXplore t-SNE overview figure with superimposed classification information. A) K-medoid cluster numbers zero to seven of a clustering with k set to eight are all highlighted in color, exhausting the provided color scale. Clusters provide complementary insights into what can be considered roughly proximal in the t-SNE projection. B) Making use of ms2query analog classifications as proxies for chemical classification we can see that analog matches for our spectra already show consistent chemical organization in the data (using NPClassifier pathway ontology).

When running k-medoid clustering with larger values of k, the feature groups will tend to become smaller and align well with t-SNE positioning and topological connectivity patterns. Inside specXplore, this can be seen via groupings tending to contain only feature nodes in close proximity to one another in the t-SNE embedding 3.1.4.

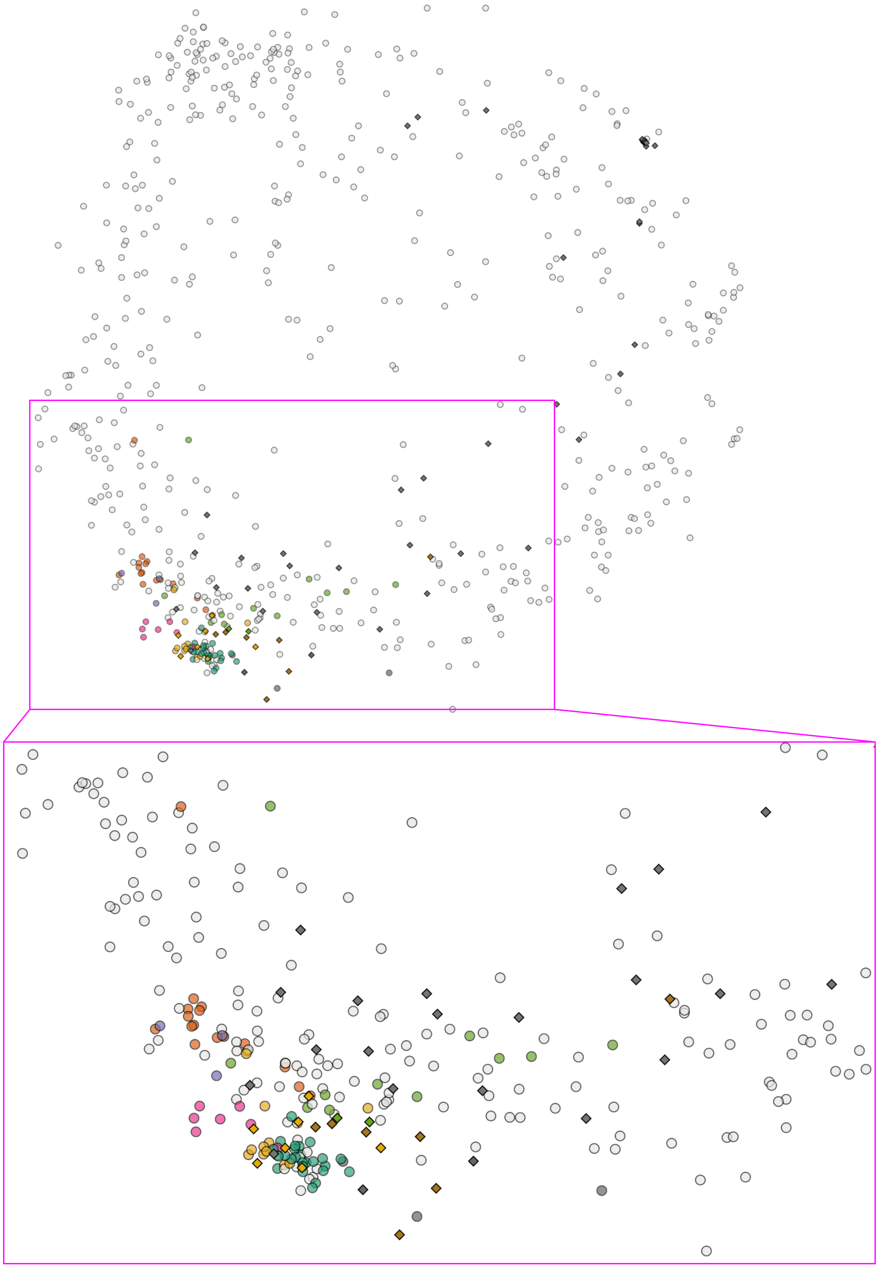
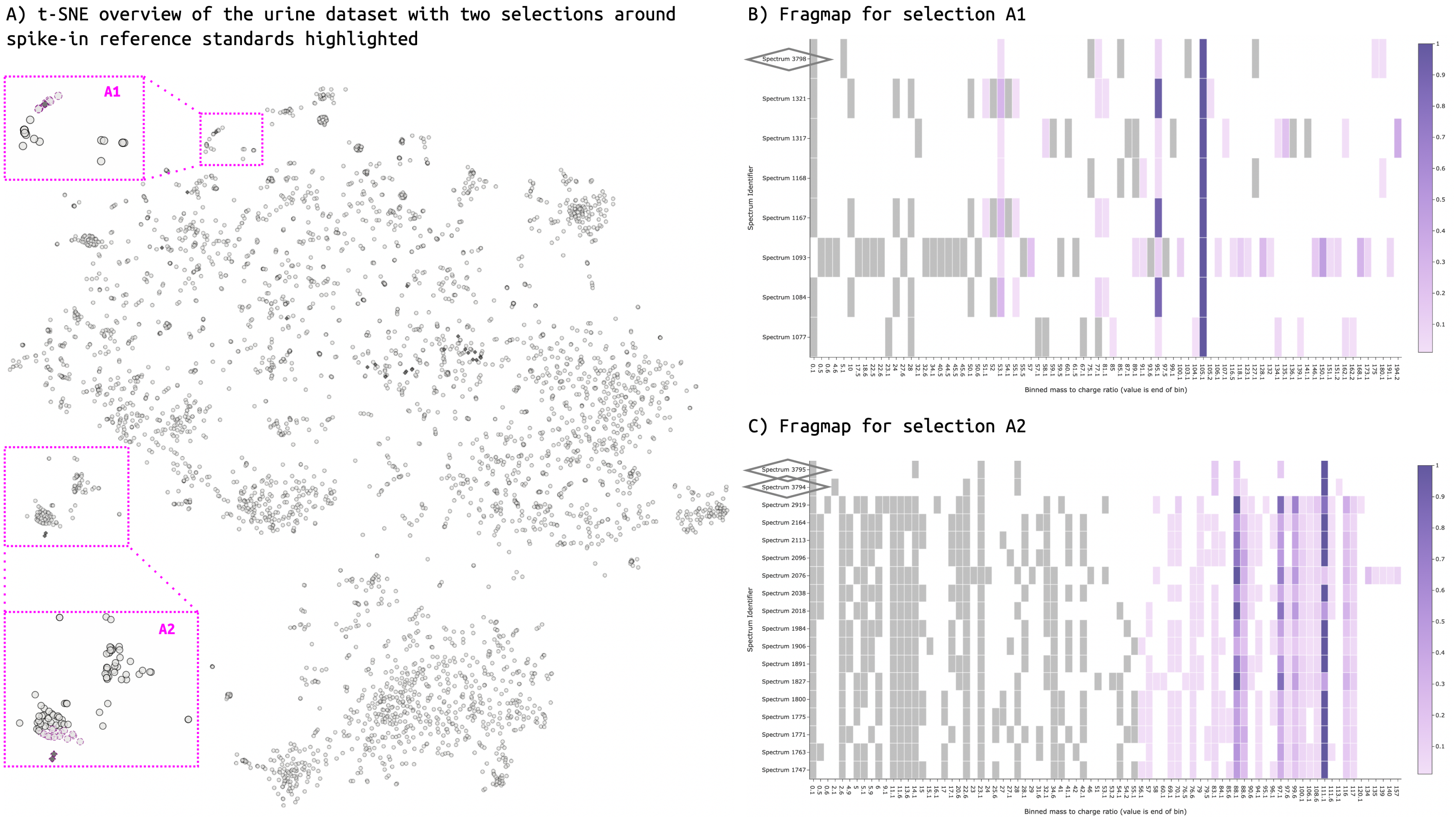


Figure 4: k-medoid clusters in lower left area of the wheat data set example at k = 120. While some node groups are more densely arranged in the t-SNE embedding than others, most groups tend to be highly localized in one area of the embedding.

3.2 Urine Data Example

In addition to the wheat data example shown in the main manuscript, we have also run specXplore on an urine exposomics dataset [15]. Here, study participants were subjected to low and high polyphenol dietary interventions to study the impact of high polyphenol diets on the urine metabolome. Urine samples were collected in three distinct phases. First, subjects were put on low polyphenol diet for three days, and their urine samples were sampled for 24 hours of the third day. Second, immediately following the three day consecutive low polyphenol diet, a polyphenol rich smoothie was ingested in the morning of day four. Urine was sampled throughout this fourth day while participants continued the otherwise low polyphenol diet. Third, on the day after the high polyphenol smoothie, subjects were consuming a high polyphenol diet of their choosing and urine samples were measured throughout the day. Samples were measured with high-resolution mass spectrometry (HRMS) and processed with MZMine3 [15]. This dataset is substantially larger than the wheat dataset, containing 3818 spectra after specXplore filtering steps. However, as can be seen in Figure 3.2, local explorations in the larger t-SNE overview still work similarly. Local node groups show fragmentation overlaps and spectral similarity that can be used to find related groups of spectra. Two examples of small selections of features are shown around spike-in reference standards for illustrative purposes. In both cases, fragmentation overview maps show good overlap in characteristic fragment ions cross the selected nodes.


A) t-SNE overview of the urine dataset with two selections around spike-in reference standards highlighted

B) Fragmap for selection A1

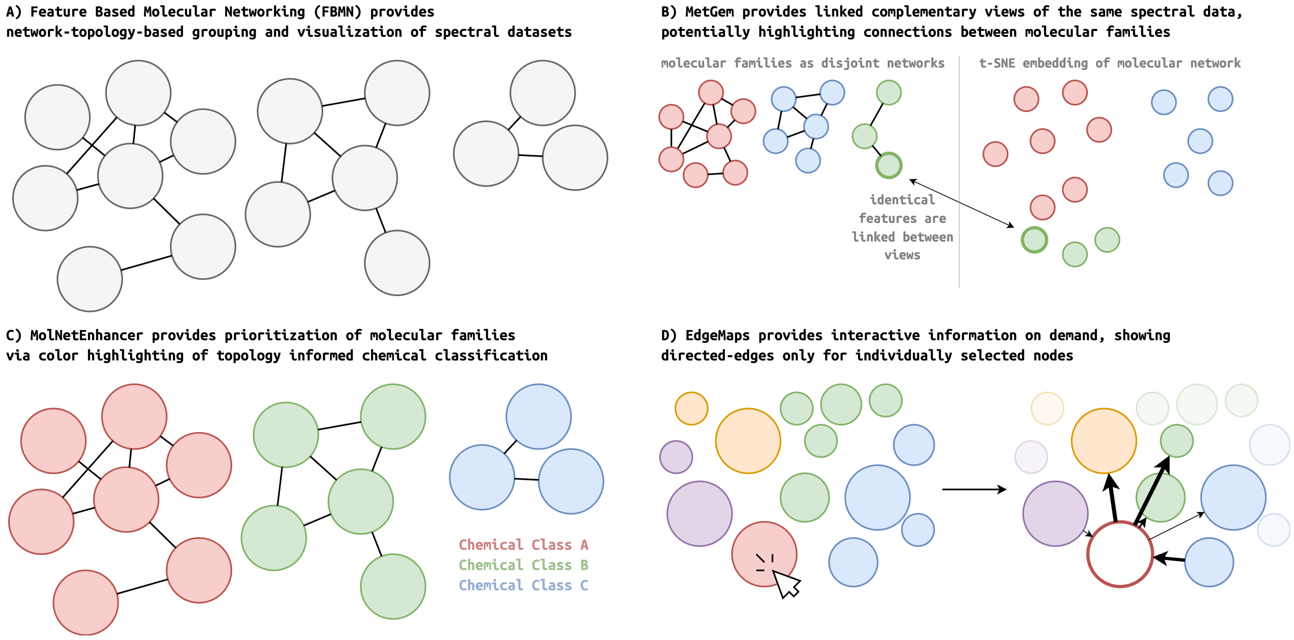
C) Fragmap for selection A2

Figure 5: Urine data exploration. A) t-SNE overview representation for the urine dataset. Experimental spectra are represented by circular grey nodes while in-silico spike-in reference standards are highlighted as darker diamond nodes. Two small local node environments involving reference spectra are highlighted. These are in close proximity in the t-SNE embedding and have substantial fragmentation overlap. B) Fragmap for selection A1. There are many fragmentation differences among the spectra. However, high relative intensity fragment ion bins are shared across all spectra in the selection. Spike-in reference spectrum additionally highlighted using diamond outline on fragmap y axis. C) Fragmap for selection A2. Large overlap in binned fragmentation pattern exists between all spectra in this selection. This includes higher intensity fragments, lower intensity fragments, and neutral losses (gray).

4 Dashboard

4.1 Graphical Overview of Methods that inspired specXplore

Figure 6 graphically represents molecular FBMN (Feature Based Molecular Networking), MetGem, MolNetEnhancer, and EdgeMaps. All four of these tools inspired design choices made in specXplore.



A) Feature Based Molecular Networking (FBMN) provides network-topology-based grouping and visualization of spectral datasets

B) MetGem provides linked complementary views of the same spectral data, potentially highlighting connections between molecular families

C) MolNetEnhancer provides prioritization of molecular families via color highlighting of topology informed chemical classification

D) EdgeMaps provides interactive information on demand, showing directed-edges only for individually selected nodes

molecular families as disjoint networks
t-SNE embedding of molecular network
identical features are linked between views
Chemical Class A
Chemical Class B
Chemical Class C

Figure 6: SpecXplore is an exploratory dashboard aimed at exploring MS/MS data generated from LC-MS/MS datasets. It is inspired by and related to A) FBMN (Feature Based Molecular Networking) through the premise of molecular networking [16], B) MetGem through the use of a low dimensional embedding in a complementary fashion alongside network views [6], C) MolNetEnhancer (Molecular Network Enhancer) through the idea of joining color and classifications to help users prioritize sub-networks of interest [17], and D) EdgeMaps through the approach to unlocking details including edges in the network views in an interactive fashion [18]. All illustrations are conceptual simplifications and do not represent the exact visual choices of the tools in the respective papers.

4.2 MS/MS data comparison views: fragmap, fragmentation spectra, and mirror plots

SpecXplore makes use of dynamic thresholds in a graph exploration setting. It is hence important to provide the user with insights into whether increased connectivity as a consequence of more liberal thresholds is still reliable or useful. To assist with this, a fragmentation overview heatmap inspired by *UpSet* plots [19] is provided that allows visualizing overlaps in fragmentation between sets of spectra. This fragmentation map, hereafter called fragmap, provides a generalization of a mirror plot but also improves upon the visual use of the space available for plotting. This is done by employing a factorized and sorted recasting of binned mass-to-charge ratios. Here, factorized mass-to-charge ratios are put onto the x-axis in order, while the y-axis delineates the spectra. For each mass-to-charge ratio bin, intensities of the fragment in the respective spectra are displayed using a color gradient as typical in heatmaps. Fragment overlap can be assessed instantly for each fragment between spectra, while intensity agreements or disagreements are easily spotted. Importantly, the x-axis is the same for all spectra and the plot loses minimal plotting surface to white spaces as distance in mass to charge ratios between fragments is removed through factorization. Optionally, each spectrum is filtered down to its respective top-K highest intensity fragments before factorization and neutral loss computation to provide a simplified view for complex or noisy spectra. The result is a concise overview of fragmentation overlaps between spectra for quick assessments of relatedness between matched nodes. While fragmaps provide a means of quickly asserting overlaps across spectra, they perform a number of transformation steps that can obscure the raw data. Most notably, binning may aggregate different fragment peaks into the same bin, leading to loss of information and possibly misleading overlaps in the fragmentation overview. This potential bias is counteracted using interactive hover pop-ups that provide the original mass-to-charge ratio values and intensities of any fragments being aggregated into a single bin, for each bin-fragment and spectrum separately. The user can thus quickly assess the bin artefacts when inspecting promising overlaps. In addition, to provide a more direct and familiar link to the raw spectra themselves, specXpolore also provides spectrum plots and mirror plots.

4.3 Color as a medium for prioritization and visual emphasis

In general, color is one of the most effective visual channels available to us; both for categorical and continuous data. In specXplore, we use color as a tool for emphasis and guiding attention. For instance, spike-in standards are highlighted in darker gray colors and increased opacity compared to experimental nodes. Any selected nodes are encircled with magenta highlighting. In egonets, we make use of the Viridis color scheme discretized to n colors, where n is the number of hops allowed, to indicate distance from the egonet. Here, color is further reinforced by opacity and line thickness changes as a function of hop distance. In addition, network views for selections of nodes emphasize the selected nodes and their interconnections in magenta while connections out of the selected set are visualized in black. Chemical classes or k-medoids groupings can further be highlighted in color within the t-SNE overview graph. While color provides a powerful medium for emphasis, care must be taken to avoid "color saturation", as humans are generally only able to meaningfully differentiate up to eight to twelve unique colors [20]. Additionally, color-blind-friendly color-maps should be used as much as possible [21, 22]. In specXplore, we make use of the *ColorBrewer* [23] Dark2 to provide an appropriate color palette for highlighting up to eight different node classes in the overview t-SNE graph. In addition, to make our augmented heatmap representation more useful to colorblind users, we incorporated a toggle for grayscale similarity depiction with magenta markers, thus avoiding combinations of blue, red, and green at the same time.

4.4 Chemical Class Visualization

Chemical ontologies provide a scientific foundation for data subdivision [24, 25]. However, ontologies require known chemical structures which are generally not available for experimental mass spectral data explored using molecular networking. Different means of providing chemical information for unknown spectra exist such as Canopus for chemical classification [11, 12] or ms2lda for chemical motif detection [26]. In specXplore, the exact use of predicted ontology is left to the user and easy to integrate via feature identifier matching. The ms2query best-scoring analog classifications may be used to get a rough feeling for chemical class via putative analog class distribution in the overview graph. A current limitation in specXplore is that, at each ontology level, each feature may only be member of a single class. In practice, this means that any multiple classifications are amalgamated together into a single entry for each feature. This may lead to some loss of class select-ability for highlighting.

4.5 Effective visual analysis using scented widgets

Effective visual analysis depends on both the selected visual encoding and the means for interactively traversing said encoding. A host of possible interaction techniques present themselves to make this navigation as effective as possible [27, 28]. For specXplore, localized on-demand selections and interactive filtering of edges between nodes is used for allowing users to explore local graph topology without excessive visual clutter. Determining what threshold to use in graph visualization is a non-trivial task [29]. Inspired by Willet et al.'s *Scented Widgets* [30], specXplore provides a visual summary of the edge weight distribution to the input slider in order to provide insights into global impacts of thresholds on edge numbers. For more local impact evaluations, node degrees can be visually highlighted using a color gradient from the minimum node size to the maximum node size in the t-SNE overview. Both of these options, as well as the augmap-based matrix representation, provide insights into the topological impacts of thresholds and thus allow exploration in an informed manner rather than random trial and error. Of course, metabolomics domain considerations will ultimately determine the acceptance or rejection of edges.

4.6 Augmented Matrix Views

We make use of a multilevel heatmap representation of the pairwise similarity matrix to facilitate quantitative insight generation into the actual ms2deepscore pairwise similarity matrix, but also to allow comparison between scores from the ms2deepscore anchor point. Different similarity measures for the same features can be viewed together as a multi-layer graph, i.e. a set of nodes connected by different "layers" of edges (in this case) defined by each similarity measure [31]. Visualizing such graphs is non-trivial, and a number of visualization techniques have been put forth [32]. In specXplore, our graph comparison aims to imply that we are interested in visualizing all the layers of the multi-graph simultaneously. Given the visual clutter that rendering all such edges would introduce, a conventional network representation would be rendered nigh unreadable. Similarly, radial embeddings [33] as well as so-called *Hive Plots* [34] were considered, but ultimately deemed inappropriate. Instead, we opt for a matrix representation, as they have been shown to be more effective and preferred (in all but path-tracing tasks [35]) when applied to large and/or dense graphs [36, 37]. Importantly, when using the matrix representation in the form of a heatmap, quantitative information can be color encoded for one of the measures, with additional markers giving qualitative adjacency insights into the other metrics. Such marker-augmented heatmaps are uncommon, though there is precedent in other fields (e.g. figure 5 in [38]).

4.7 Tabular Metadata Integration

Not all metadata lends itself well to graphical visualization in the spectral exploration dashboard, but may nonetheless be useful to have integrated into the tool for quick access upon node selection. SpecXplore offers a metadata table panel that can be generated for any node selection, allowing the user to inspect any additional supplied information regarding a spectrum inside the dashboard. This is where ms2query analog hits, but also any user-provided metadata can be displayed for selections of nodes. Incorporation of this table allows users to explore their data within specXplore and thus removes some of the need for switching between programs. That being said, it also allows users to inspect relevant identifiers in the metadata to quickly inspect features in other programs as well.

5 Bibliography

[1] Robin Schmid, Steffen Heuckeroth, Ansgar Korf, Aleksandr Smirnov, Owen Myers, Thomas S. Dyrlund, Roman Bushuiev, Kevin J. Murray, Nils Hoffmann, Miaoshan Lu, Abinesh Sarvepalli, Zheng Zhang, Markus Fleischauer, Kai Dührkop, Mark Wesner, Shawn J. Hoogstra, Edward Rudt, Olena Mokshyna, Corinna Brungs, Kirill Ponomarov, Lana Mutabdžija, Tito Damiani, Chris J. Pudney, Mark Earll, Patrick O. Helmer, Timothy R. Fallon, Tobias Schulze, Albert Rivas-Ubach, Aivett Bilbao, Henning Richter, Louis-Félix Nothias, Mingxun Wang, Matej Orešič, Jing-Ke Weng, Sebastian Böcker, Astrid Jeibmann, Heiko Hayen, Uwe Karst, Pieter C. Dorrestein, Daniel Petras, Xiuxia Du, and Tomáš Pluskal. Integrative analysis of multimodal mass spectrometry data in MZmine 3. *Nature Biotechnology*, 41(4):447–449, March 2023. doi: 10.1038/s41587-023-01690-2. URL https://doi.org/10.1038/s41587-023-01690-2.

[2] Angelos Chatzimparmpas, Rafael M. Martins, and Andreas Kerren. t-visne: Interactive assessment and interpretation of t-sne projections. *IEEE Transactions on Visualization and Computer Graphics*, 26(8): 2696–2714, 2020. doi: 10.1109/TVCG.2020.2986996.

[3] Daniel Probst and Jean-Louis Reymond. Visualization of very large high-dimensional data sets as minimum spanning trees. *Journal of Cheminformatics*, 12(1):12, Feb 2020. ISSN 1758-2946. doi: 10.1186/s13321-020-0416-x. URL https://doi.org/10.1186/s13321-020-0416-x.

[4] Pierre-Marie Allard, Arnaud Gaudry, Luis-Manuel Quirós-Guerrero, Adriano Rutz, Miwa Dounoue-Kubo, Tom W N Walker, Emmanuel Defossez, Christophe Long, Antonio Grondin, Bruno David, and Jean-Luc Wolfender. Open and reusable annotated mass spectrometry dataset of a chemodiverse collection of 1,600 plant extracts. *GigaScience*, 12, 01 2023. ISSN 2047-217X. doi: 10.1093/gigascience/giac124. URL https://doi.org/10.1093/gigascience/giac124. giac124.

[5] Yingfan Wang, Haiyang Huang, Cynthia Rudin, and Yaron Shaposhnik. Understanding how dimension reduction tools work: an empirical approach to deciphering t-sne, umap, trimap, and pacmap for data visualization. *The Journal of Machine Learning Research*, 22(1):9129–9201, 2021.

[6] Florent Olivon, Nicolas Elie, Gwendal Grelier, Fanny Roussi, Marc Litaudon, and David Touboul. Metgem software for the generation of molecular networks based on the t-sne algorithm. *Analytical Chemistry*, 90 (23):13900–13908, 2018. doi: 10.1021/acs.analchem.8b03099. URL https://doi.org/10.1021/acs.analchem.8b03099. PMID: 30335965.

[7] Florian Huber, Sven van der Burg, Justin J. J. van der Hooft, and Lars Ridder. Ms2deepscore: a novel deep learning similarity measure to compare tandem mass spectra. *Journal of Cheminformatics*, 13(1): 84, Oct 2021. ISSN 1758-2946. doi: 10.1186/s13321-021-00558-4. URL https://doi.org/10.1186/s13321-021-00558-4.

[8] Niek F. de Jonge, Joris R. Louwen, Elena Chekmeneva, Stephane Camuzeaux, Femke J. Vermeir, Robert S. Jansen, Florian Huber, and Justin J.J. van der Hooft. Ms2query: Reliable and scalable ms2 mass spectral-based analogue search. *bioRxiv*, 2022. doi: 10.1101/2022.07.22.501125. URL https://www.biorxiv.org/content/early/2022/07/23/2022.07.22.501125.

[9] Jeramie Watrous, Patrick Roach, Theodore Alexandrov, Brandi S. Heath, Jane Y. Yang, Roland D. Kersten, Menno van der Voort, Kit Pogliano, Harald Gross, Jos M. Raaijmakers, Bradley S. Moore, Julia Laskin, Nuno Bandeira, and Pieter C. Dorrestein. Mass spectral molecular networking of living microbial colonies. *Proceedings of the National Academy of Sciences*, 109(26):E1743–E1752, 2012. doi: 10.1073/pnas.1203689109. URL https://www.pnas.org/doi/abs/10.1073/pnas.1203689109.

[10] Florian Huber, Lars Ridder, Stefan Verhoeven, Jurriaan H. Spaaks, Faruk Diblen, Simon Rogers, and Justin J. J. van der Hooft. Spec2vec: Improved mass spectral similarity scoring through learning of structural relationships. *PLOS Computational Biology*, 17(2):1–18, 02 2021. doi: 10.1371/journal.pcbi.1008724. URL https://doi.org/10.1371/journal.pcbi.1008724.

[11] Kai Dührkop, Markus Fleischauer, Marcus Ludwig, Alexander A. Aksenov, Alexey V. Melnik, Marvin Meusel, Pieter C. Dorrestein, Juho Rousu, and Sebastian Böcker. Sirius 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nature Methods*, 16(4):299–302, Apr 2019. ISSN 1548-7105. doi: 10.1038/s41592-019-0344-8. URL https://doi.org/10.1038/s41592-019-0344-8.

[12] Kai Dührkop, Louis-Félix Nothias, Markus Fleischauer, Raphael Reher, Marcus Ludwig, Martin A Hoffmann, Daniel Petras, William H Gerwick, Juho Rousu, Pieter C Dorrestein, and Sebastian Böcker. Systematic classification of unknown metabolites using high-resolution fragmentation mass spectra. *Nat Biotechnol*, 39(4):462–471, November 2020.

[13] Christoph Bueschl, Maria Doppler, Elisabeth Varga, Bernhard Seidl, Mira Flasch, Benedikt Warth, and Juergen Zanghellini. Peakbot: machine-learning-based chromatographic peak picking. *Bioinformatics*, 38 (13):3422–3428, 2022.

[14] Maria Doppler, Christoph Bueschl, Bernhard Kluger, Andrea Koutnik, Marc Lemmens, Hermann Buerstmayr, Justyna Rechthaler, Rudolf Krska, Gerhard Adam, and Rainer Schuhmacher. Stable isotope–assisted plant metabolomics: Combination of global and tracer-based labeling for enhanced untargeted profiling and compound annotation. *Frontiers in Plant Science*, 10:1366, 2019.

[15] Ian Oesterle, Manuel Pristner, Sabrina Berger, Mingxun Wang, Vinicius Verri Hernandes, Annette Rompel, and Benedikt Warth. Exposomic biomonitoring of polyphenols by non-targeted analysis and suspect screening. *Analytical Chemistry*, 95(28):10686–10694, 2023.

[16] Louis-Félix Nothias, Daniel Petras, Robin Schmid, Kai Dührkop, Johannes Rainer, Abinesh Sarvepalli, Ivan Protsyuk, Madeleine Ernst, Hiroshi Tsugawa, Markus Fleischauer, Fabian Aicheler, Alexander A. Aksenov, Oliver Alka, Pierre-Marie Allard, Aiko Barsch, Xavier Cachet, Andres Mauricio Caraballo-Rodriguez, Ricardo R. Da Silva, Tam Dang, Neha Garg, Julia M. Gauglitz, Alexey Gurevich, Giorgis Isaac, Alan K. Jarmusch, Zdeněk Kameník, Kyo Bin Kang, Nikolas Kessler, Irina Koester, Ansgar Korf, Audrey Le Gouellec, Marcus Ludwig, Christian Martin H., Laura-Isobel McCall, Jonathan McSayles, Sven W. Meyer, Hosein Mohimani, Mustafa Morsy, Oriane Moyne, Steffen Neumann, Heiko Neuweger, Ngoc Hung Nguyen, Melissa Nothias-Esposito, Julien Paolini, Vanessa V. Phelan, Tomáš Pluskal, Robert A. Quinn, Simon Rogers, Bindesh Shrestha, Anupriya Tripathi, Justin J. J. van der Hooft, Fernando Vargas, Kelly C. Weldon, Michael Witting, Heejung Yang, Zheng Zhang, Florian Zubeil, Oliver Kohlbacher, Sebastian Böcker, Theodore Alexandrov, Nuno Bandeira, Mingxun Wang, and Pieter C. Dorrestein. Feature-based molecular networking in the gnps analysis environment. *Nature Methods*, 17(9):905–908, Sep 2020. ISSN 1548-7105. doi: 10.1038/s41592-020-0933-6. URL https://doi.org/10.1038/s41592-020-0933-6.

[17] Madeleine Ernst, Kyo Bin Kang, Andrés Mauricio Caraballo-Rodríguez, Louis-Felix Nothias, Joe Wandy, Christopher Chen, Mingxun Wang, Simon Rogers, Marnix H. Medema, Pieter C. Dorrestein, and Justin J.J. van der Hooft. Molnetenhancer: Enhanced molecular networks by integrating metabolome mining and annotation tools. *Metabolites*, 9(7), 2019. ISSN 2218-1989. doi: 10.3390/metabo9070144. URL https://www.mdpi.com/2218-1989/9/7/144.

[18] Marian Dörk, Sheelagh Carpendale, and Carey Williamson. EdgeMaps: visualizing explicit and implicit relations. In *Visualization and Data Analysis 2011*. SPIE, January 2011. doi: 10.1117/12.872578. URL https://doi.org/10.1117/12.872578.

[19] Alexander Lex, Nils Gehlenborg, Hendrik Strobelt, Romain Vuillemot, and Hanspeter Pfister. UpSet: Visualization of Intersecting Sets. *IEEE Transactions on Visualization and Computer Graphics*, 20(12): 1983–1992, December 2014. ISSN 1941-0506. doi: 10.1109/TVCG.2014.2346248. Conference Name: IEEE Transactions on Visualization and Computer Graphics.

[20] Tamara Munzner. *Visualization Analysis and Design*. A K Peters/CRC Press, New York, October 2014. ISBN 978-0-429-08890-2. doi: 10.1201/b17511.

[21] Luke Jefferson and Richard Harvey. Accommodating color blind computer users. In *Proceedings of the 8th international ACM SIGACCESS conference on Computers and accessibility*, Assets '06, pages 40–47, New York, NY, USA, October 2006. Association for Computing Machinery. ISBN 978-1-59593-290-7. doi: 10.1145/1168987.1168996. URL https://doi.org/10.1145/1168987.1168996.

[22] Frank Elavsky, Cynthia Bennett, and Dominik Moritz. How accessible is my visualization? Evaluating visualization accessibility with Chartability. *Computer Graphics Forum*, 41(3):57–70, 2022. ISSN 1467-8659. doi: 10.1111/cgf.14522. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.14522. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.14522.

[23] Mark Harrower and Cynthia A. Brewer. ColorBrewer.org: An Online Tool for Selecting Colour Schemes for Maps. *The Cartographic Journal*, 40(1):27–37, June 2003. ISSN 0008-7041. doi: 10.1179/000870403235002042. URL https://www.tandfonline.com/doi/abs/10.1179/000870403235002042. Publisher: Taylor & Francis _eprint: https://www.tandfonline.com/doi/pdf/10.1179/000870403235002042.

[24] Yannick Djoumbou Feunang, Roman Eisner, Craig Knox, Leonid Chepelev, Janna Hastings, Gareth Owen, Eoin Fahy, Christoph Steinbeck, Shankar Subramanian, Evan Bolton, Russell Greiner, and David S. Wishart. Classyfire: automated chemical classification with a comprehensive, computable taxonomy. *Journal of Cheminformatics*, 8(1):61, Nov 2016. ISSN 1758-2946. doi: 10.1186/s13321-016-0174-y. URL https://doi.org/10.1186/s13321-016-0174-y.

[25] Hyun Woo Kim, Mingxun Wang, Christopher A. Leber, Louis-Félix Nothias, Raphael Reher, Kyo Bin Kang, Justin J. J. van der Hooft, Pieter C. Dorrestein, William H. Gerwick, and Garrison W. Cottrell. Npclassifier: A deep neural network-based structural classification tool for natural products. *Journal of Natural Products*, 84(11):2795–2807, 2021. doi: 10.1021/acs.jnatprod.1c00399. URL https://doi.org/10.1021/acs.jnatprod.1c00399. PMID: 34662515.

[26] Justin Johan Jozias van der Hooft, Joe Wandy, Michael P. Barrett, Karl E. V. Burgess, and Simon Rogers. Topic modeling for untargeted substructure exploration in metabolomics. *Proceedings of the National Academy of Sciences*, 113(48):13738–13743, 2016. doi: 10.1073/pnas.1608041113. URL https://www.pnas.org/doi/abs/10.1073/pnas.1608041113.

[27] Ji Soo Yi, Youn ah Kang, John Stasko, and J.A. Jacko. Toward a Deeper Understanding of the Role of Interaction in Information Visualization. *IEEE Transactions on Visualization and Computer Graphics*, 13 (6):1224–1231, November 2007. ISSN 1941-0506. doi: 10/fmrs6r. Conference Name: IEEE Transactions on Visualization and Computer Graphics.

[28] Ana Figueiras. Towards the Understanding of Interaction in Information Visualization. In *2015 19th International Conference on Information Visualisation*, pages 140–147, July 2015. doi: 10.1109/iV.2015.34. ISSN: 2375-0138.

[29] Jinwook Seo and B. Shneiderman. A Rank-by-Feature Framework for Unsupervised Multidimensional Data Exploration Using Low Dimensional Projections. In *IEEE Symposium on Information Visualization*, pages 65–72, October 2004. doi: 10.1109/INFVIS.2004.3. ISSN: 1522-404X.

[30] Wesley Willett, Jeffrey Heer, and Maneesh Agrawala. Scented Widgets: Improving Navigation Cues with Embedded Visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1129–1136, November 2007. ISSN 1941-0506. doi: 10.1109/TVCG.2007.70589. Conference Name: IEEE Transactions on Visualization and Computer Graphics.

[31] Mikko Kivelä, Alex Arenas, Marc Barthelemy, James P. Gleeson, Yamir Moreno, and Mason A. Porter. Multilayer networks. *Journal of Complex Networks*, 2(3):203–271, September 2014. ISSN 2051-1310. doi: 10.1093/comnet/cnu016. URL https://doi.org/10.1093/comnet/cnu016.

[32] F. McGee, M. Ghoniem, G. Melançon, B. Otjacques, and B. Pinaud. The State of the Art in Multilayer Network Visualization. *Computer Graphics Forum*, 38(6):125–149, 2019. ISSN 1467-8659. doi: 10.1111/cgf.13610. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.13610. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.13610.

[33] Martin Krzywinski, Jacqueline Schein, İnanç Birol, Joseph Connors, Randy Gascoyne, Doug Horsman, Steven J. Jones, and Marco A. Marra. Circos: An information aesthetic for comparative genomics. *Genome Research*, 19(9):1639–1645, September 2009. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.092759.109. URL https://genome.cshlp.org/content/19/9/1639. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

[34] Martin Krzywinski, Inanc Birol, Steven JM Jones, and Marco A Marra. Hive plots—rational approach to visualizing networks. *Briefings in Bioinformatics*, 13(5):627–644, September 2012. ISSN 1467-5463. doi: 10.1093/bib/bbr069. URL https://doi.org/10.1093/bib/bbr069.

[35] Mohammad Ghoniem, Jean-Daniel Fekete, and Philippe Castagliola. On the Readability of Graphs Using Node-Link and Matrix-Based Representations: A Controlled Experiment and Statistical Analysis. *Information Visualization*, 4(2):114–135, June 2005. ISSN 1473-8716, 1473-8724. doi: 10.1057/palgrave.ivs.9500092. URL http://journals.sagepub.com/doi/10.1057/palgrave.ivs.9500092.

[36] René Keller, Claudia M. Eckert, and P. John Clarkson. Matrices or Node-Link Diagrams: Which Visual Representation is Better for Visualising Connectivity Models? *Information Visualization*, 5(1):62–76, March 2006. ISSN 1473-8716, 1473-8724. doi: 10.1057/palgrave.ivs.9500116. URL http://journals.sagepub.com/doi/10.1057/palgrave.ivs.9500116.

[37] M. Ghoniem, J.-D. Fekete, and P. Castagliola. A Comparison of the Readability of Graphs Using Node-Link and Matrix-Based Representations. In *IEEE Symposium on Information Visualization*, pages 17–24, October 2004. doi: 10.1109/INFVIS.2004.1. ISSN: 1522-404X.

[38] Bumhee Park, Dae-Shik Kim, and Hae-Jeong Park. Graph independent component analysis reveals repertoires of intrinsic network components in the human brain. *PLOS ONE*, 9(1):1–10, 01 2014. doi: 10.1371/journal.pone.0082873. URL https://doi.org/10.1371/journal.pone.0082873.