Title

Network Topology Evaluation and Transitive Alignments for Molecular Networking.

Permalink

https://escholarship.org/uc/item/4bq907gr

Journal

Journal of the American Society for Mass Spectrometry, 35(9)

Authors

Wang, Xianghu
Strobel, Michael
Aron, Allegra
et al.

Publication Date

2024-09-04

DOI

10.1021/jasms.4c00208

Published in final edited form as:
J Am Soc Mass Spectrom. 2024 September 04; 35(9): 2165–2175. doi:10.1021/jasms.4c00208.

Network Topology Evaluation and Transitive Alignments for Molecular Networking

Xianghu Wang[1], **Michael Strobel**[1], **Allegra T Aron**[2], **Vanessa V Phelan**[3], **Deepa D Acharya**[4], **Ken Clevenger**[4], **Christopher J Brown**[5], **Jie Hu**[6], **Ashley Kretsch**[4], **Elizabeth H Mahood**[6], **Carla Menegatti**[4], **Quanbo Xiong**[4], **Mingxun Wang**[1]

[1]Department of Computer Science and Engineering, University of California Riverside, 900 University Ave., Riverside, CA, 92521, United States

[2]Department of Chemistry and Biochemistry, University of Denver, 2101 East Wesley Ave, Denver, CO, 80210, United States

[3]Department of Pharmaceutical Sciences, Skaggs School of Pharmacy and Pharmaceutical Sciences, University of Colorado, Anschutz Medical Campus, 12850 E Montview Blvd, Aurora, CO, 80045, United States, United States

[4]Biologicals Research and Development, Corteva Agriscience, 9330 Zionsville Rd, Indianapolis, IN 46268

[5]Regulatory Science, Corteva Agriscience, 9330 Zionsville Rd, Indianapolis, IN 46268

[6]Data Science, Corteva Agriscience, 9330 Zionsville Rd, Indianapolis, IN 46268

Abstract

Untargeted tandem mass spectrometry (MS/MS) is an essential technique in modern analytical chemistry, providing a comprehensive snapshot of chemical entities in complex samples and identifying unknowns through their fragmentation patterns. This high-throughput approach generates large datasets that can be challenging to interpret. Molecular Networks (MNs) have been developed as a computational tool to aid in the organization and visualization of complex chemical space in untargeted mass spectrometry data, thereby supporting comprehensive data analysis and interpretation. MNs group related compounds with potentially similar structures from MS/MS data

Author Contributions
MW conceptualized the idea. MW organized the project. XW, MW, QX wrote the manuscript. XW and MW developed the software. All authors approve of the manuscript.

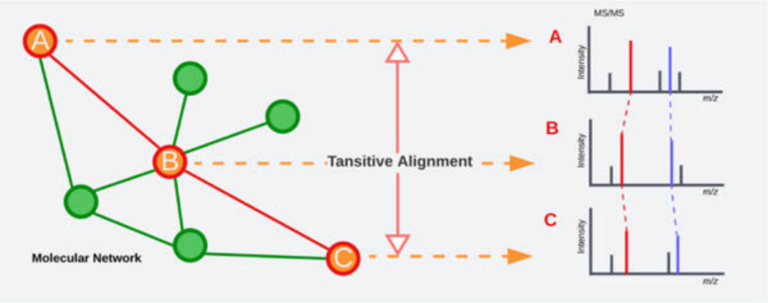
Source Code Availability
Benchmarking repository:
https://github.com/XianghuWang-287/MN_benchmakring
Zenodo: https://zenodo.org/records/10729443
Transitive Alignment workflow repository:
https://github.com/Wang-Bioinformatics-Lab/Transitive_alignment_workflow
Zenodo: https://zenodo.org/records/10779473

Supporting Information
Molecular Network benchmarking results for all the testing datasets, Top K parameter choosing evaluation, network sparsity analysis results, detailed information about the benchmarking results, and transitive alignment supporting information for the Corteva dataset (DOC)

Competing interest statement
MW is a cofounder of Ometa Labs LLC

by calculating all pairwise MS/MS similarities and filtering these connections to produce a MN. Such networks are instrumental in metabolomics for identifying novel metabolites, elucidating metabolic pathways, and even discovering biomarkers for disease. While MS/MS similarity metrics have been explored in the literature, the influence of network topology approaches on MN construction remains unexplored. This manuscript introduces metrics for evaluating MN construction, benchmarks state-of-the-art approaches, and proposes the Transitive Alignments approach to improve MN construction. The Transitive Alignment technique leverages the MN topology to re-align MS/MS spectra of related compounds that differ by multiple structural modifications. Combining this Transitive Alignments approach with pseudo-clique finding, a method for identifying highly-connected groups of nodes in a network, resulted in more complete and higher quality molecular families. Finally, we also introduce a targeted network construction technique called induced transitive alignments where we demonstrate effectiveness on a real world natural product discovery application. We release this transitive alignment technique as a high-throughput workflow that can be used by the wider research community.

Graphical Abstract



The core of Transitive Alignment is to use the tandem mass spectrum of the intermedia nodes to re-align the spectra of two indirectly connected nodes within the molecular network

Introduction

Mass spectrometry is a key analytical technique for the characterization of complex small molecule biological samples[1]. Specifically, tandem mass spectrometry (MS/MS) is used extensively to measure small molecules by detecting the pattern of fragmentation of these molecules[2]. Modern mass spectrometry instrumentation can collect tens of thousands of MS/MS spectra in a single analysis in minutes. The ability to interpret these MS/MS spectra, i.e. annotate the chemical compounds that the measurements represent, or prioritize for downstream analysis are key challenges that can unlock new biological discoveries[3–5].

Molecular Networking (MN) is a computational approach that aims to organize and visualize the diverse chemistry observed in untargeted mass spectrometry data. At its core, MNs aim to group different, yet structurally similar molecules together, using tandem mass spectrometry (MS/MS) data. The MN approach utilizes alignments between every pair of MS/MS spectra to achieve this, accounting for structural modifications between the molecules[6–9]. This alignment is referred to either as an aligned cosine or modified cosine method. The two distinct phases of MN construction are: 1) the full pairwise comparison of all MS/MS in a dataset with the aligned cosine method, and 2) a pruning of these comparisons, which involves selectively removing potentially incorrect or redundant connections, to simplify the molecular network for analysis and visualization by chemists. While this first step has received significant research attention, the second (called topology filtering) has been neglected. The current state of the art of topology filtering employs heuristic methods that have not been rigorously evaluated[6,7]. Here, we address these shortcomings by introducing a method to evaluate the construction of MNs by introducing metrics that balance the completeness and accuracy of MNs.

Given the findings in our benchmarking, we identified key shortcomings in the state-of-the-art topology filtering methods[10–12]. Specifically, topology filtration methods do not adequately complement a key weakness of the modified cosine method, i.e., modified cosine cannot align MS/MS spectra between compounds that differ by two or more modifications. This limitation results in small and potentially less inaccurate MNs. To overcome this challenge, we introduce a method called Transitive Alignments. The intuition of Transitive Alignments utilizes the network topology information to bridge two molecules that are not directly connected in the MN and may have multiple modifications at different sites. Our evaluation demonstrates that Transitive Alignments enhance the accuracy and completeness of MN construction.

Results

Evaluation of Network Topology Algorithms

We introduce two key metrics for evaluating the completeness and accuracy of MNs: the Network Accuracy Score and Molecular Network N20 (See Methods Network Accuracy Metric and Network Component Size Metric). Briefly, the Network Accuracy score measures the correctness (measured by Tanimoto structural similarity of the true 2D chemical structures) of the edges between MS/MS spectra in the filtered MN. The Molecular Network N20, in a single number is a measure of how large the components are in a MN. Specifically Molecular Network N20 measures the size of the smallest MN connected component, where the total number of nodes in larger components covers at least 20% of the unique MS/MS in the full molecular network. We benchmark, using these two metrics, a baseline unfiltered network and a commonly used heuristic approach, called the GNPS classic method[7]. Across a broad range of hyperparameters, we found that the GNPS classic method outperforms an unfiltered baseline (Figure 2). While an improvement over the unfiltered baseline, the default setting of the classic method in GNPS may not yield optimal results (See in Figure 2, the orange star represents the default parameter setting). There exist several other points that lie above and to the right, signifying more correct and larger MNs, of the default parameters. These correspond to other combinations of hyperparameters and subsequently offer guidance for tuning the hyperparameters of the GNPS classic method. Across several benchmark datasets that vary in complexity and size, we found that the top k neighbor parameter should be set ~15 for large network (Over 1k nodes) and ~5 for small network (less than 500 nodes) (SI Figure 2). These suggestions marginally improve the performance of MN construction by increasing N20 and Network Accuracy Score (moving to the upper right in Figure 2).

We further explored traditional graph algorithms to understand how their performance relative to the above baselines. We hypothesized that cliques might offer an opportunity to find very accurate components in the MN. We implemented and benchmarked the pseudo-clique finding algorithm: CAST[13] and found that this approach increases the Network Accuracy Score when networks are small in size (e.g., N20: <10) when compared to the GNPS Classic method (Figure 2 – Green Dots). However, the CAST method cannot achieve very large components and performance diminishes compared to the GNPS Classic method above N20 = 5. This limitation stems from the modified/aligned spectrum alignment method[10–12] which struggles to align MS/MS of molecules that differ by multiple modifications occurring at different distinct locations between the two molecules (Figure 3). When these multiple modifications occur, the aligned cosine will fail to connect these MS/MS spectra in a network, resulting in missing connections between molecules that have several modifications. These missing edges restricts the largest sizes of the cliques that can be found in the network graph (Figure 1.b).

To address this shortcoming, we introduce the Transitive Alignment method, which builds transitive consistency and reintroduces the missing edges back into the network due to multiple structure modifications. To capture these edges, Transitive Alignments enables two MS/MS spectra that differ by two or more modifications to be more fully aligned by using intermediate bridging molecules that facilitate the alignment. The key is that each of these bridging molecules only differs by a single modification to any neighbor, but together in a chain, can account for multiple modifications, with peaks aligning transitively through the bridges (see Methods   Transitive Alignment Approach). In Figure 3, a practical illustration demonstrates the Transitive Alignment approach's performance in real data. This example features molecules X, Y, and Z. Notably, molecules X and Y differ by a single modification, as do molecules Y and Z, while molecules X and Z differ by two structural modifications. Here we see that X and Y exhibit high structural and spectrum similarity (0.651 and 0.816 respectively), as do Y and Z (0.918 and 0.908 respectively). However, X and Z's exhibit inconsistent similarities (0.77 and 0.0951). This disrupts the anticipated pattern, where edges connect X to Y and Y to Z, but the transitive edge from X to Z is missing due to low spectrum similarity. Our Transitive Alignment method addresses by re-aligning the peaks from X to Z, through Y, raising the cosine score from 0.0951 to 0.872.

With Transitive Alignments introducing missing edges in the network, we re-introduce clique-finding with CAST on the resulting MN measured the performance. We found that the combination of CAST + Transitive Alignment method outperforms the baseline, GNPS Classic, and CAST in the NIH SPAC, FDA Pt2, and EMBL MCF datasets. Notably, within the NIH SPAC datasets, the CAST + Transitive Alignment shows the most significant improvement from the baseline compared to any other methods. For the NIH NP, NIH NP Rd2 positive, and PSU-MSMLS datasets, although CAST + Transitive Alignment still outperforms the CAST method, we found that CAST + Transitive Alignments struggles against the GNPS Classic method at larger N20 values.

Molecular Networking Performance vs Structure Diversity Sparsity

In benchmarking the performance of CAST + Transitive Alignments vs GNPS Classic method, we found that the network density of the underlying data modulated their relative performance. Specifically, the network density measures how many neighbors each MS/MS has in the unfiltered MN, standing in as a proxy for the underlying compound structural density. We found in general as the network density decreases (increased sparsity), we observe a drop in performance of the CAST + Transitive Alignment method relative to other methods (Figure 4, SI Figure 1). This observation is held across different datasets (Table 1), and within the same dataset. We artificially subsampled the dense datasets (FDA Pt2, NIH SPAC, EMBL MCF) to create more sparse versions of the data. We varied the subsampling from 0.5 (keeping half the nodes) to 1 (keeping all the nodes), and we observe the performance of CAST + Transitive Alignment decreases (SI Figure 3). Additionally, we also observe that the relative performance of CAST + Transitive Alignment compared to GNPS Classic decreases with lower density (SI Figure 4). Specifically, we found the N20 intersection point where CAST + Transitive Alignment and GNPS Classic become equal Network Accuracy Score reduces from 50 to 20 as the subsampling rate decreases from 1 to 0.5.

MS2DeepScore MN construction evaluation

We additionally considered an alternative approach where the modified cosine score was replaced by MS2DeepScore[14]. We retrained an MS2DeepScore model without the specific benchmarking datasets. Using this retrained MS2DeepScore model, we found that the MS2DeepScore MN outperforms the baseline unfiltered modified cosine similarity (Figure 2). Further, when the GNPS Classic method is applied, the MS2DeepScore MN yields better results in the NIH SPAC and NIH NP datasets compared to a GNPS Classic MN yet underperforms in the remaining datasets featured in our benchmarking analysis (SI Figure 1). In the NIH SPAC datasets, within the N20 range from 2 to 10, the best Network Accuracy Score achieved by cosine baseline is 0.65, whereas MS2DeepScore based MN can reach a Network Accuracy Score of 0.84 in that range. However, the performance of MS2DeepScore based MN still falls short of the CAST + Transitive Alignment approach. Throughout the N20 range that all the methods are available (from 2 to 30) the scatter plot of CAST + Transitive Alignment consistently surpasses that of the MS2DeepScore based MN (Figure 2).

Classification Consistency in MNs

We investigated whether the clusters (connected components in the MN) produced by network filtering algorithms are consistent, i.e., whether the molecules in the same component are from the same compound class. We measured this by calculating the purity[15,16] of connected components across the MN by using the “superclass” classification from ClassyFire[17,18]. We benchmarked using the datasets NIH SPAC, NIH NP, FDA PT2, and NIH NP Rd2 and found that MS2DeepScore outperformed the GNPS Classic method (a higher ratio of correctly classified clusters across all N20 ranges). However, the CAST + Transitive Alignment method outperforms both across a wide range of N20 values. In this classification consistency measure, we find that CAST performs similarly to CAST +

Transitive Alignment, however, the CAST can only be applied across a very narrow range of N20 values. While the Transitive Alignments approach extends the applicability of CAST across a wider range of N20 values (generally up to N20=~15), MS2DeepScore can achieve significantly larger MNs at the cost of reduced superclass consistency (Figure 5).

Impact of Transitive alignment in natural product discovery application

For a natural products discovery program, identifying all compounds analogous to a molecule of interest is valuable, either for complete understanding of its biosynthetic pathway or to explore the full range of in-built structure-activity-relationship (SAR). When relying on cosine similarity and GNPS classic topology, these biosynthetically related molecules can often be split across different molecular families or components. To showcase the utility of transitive alignment in terms of real-world natural product applications, we demonstrate transitive alignments on an untargeted metabolomics experiment on a fungal extract. The overarching goal in this study was to find all structural variants of porritoxin compounds to facilitate isolation for SAR from a fungal strain *Alternaria* sp., DF978Z0035. Using the GNPS Classic Topology method with default parameters, four molecules that were expected to be in the same family appeared in three distinct components (Figure 6.a). Even when more relaxed parameters were used (minimum network cosine 0.5 and minimum network matched peaks 4), these four compounds remained in three components. For these specific nodes of interest, we applied a targeted version of transitive networking, called transitive induced networking. Transitive induced networking creates a network from a specified source node, incorporating only nodes within a maximum hops distance (indicates how many structure modifications from the source node), and meeting transitive alignment score criteria (See Methods Transitive Induced Network). When we applied transitive networking on Argyrotoxin A (node 255, 3 max hops, 0.3 max transitive alignment score), all four molecules were grouped together in a single component (Figure 6.b). The transitive alignment was able to bridge all these molecules and increase their alignment scores, e.g. Argyrotoxin A (node 255, *m/z* 295.118) was connected to molecule X12535623 (node 210, *m/z* 283.154), and porritoxinol (node 260, *m/z* 297.133). Originally these nodes had a cosine similarity of 0.11 and 0.07, respectively, but transitive alignments were able to elevate the score to 0.55 and 0.57, respectively (SI Table 2 and SI Figure 6). This qualitative example demonstrates that the transitive alignment approach has created a more complete molecular family while maintaining accuracy, which can greatly improve the efficiency of analog analysis during compound isolation and structure elucidation.

Discussion and Conclusion

This manuscript introduces benchmarking metrics to evaluate the molecular network topology filtering algorithms and evaluates the state of the methods on over six public MS/MS library datasets. We show that the GNPS Classic method performs better than the unfiltered baseline. Clique-finding algorithms can improve the construction of MNs in situations with small component sizes. However, these algorithms fail when attempting to form extensive components. Given these insights, we introduce and demonstrate that Transitive Alignments enhance the creation of MNs, while also acknowledging the limitations of this method. Firstly, Transitive Alignment relies on the presence of a high density of single modifications; in their absence, creating large and consistent networks is not possible. Secondly, Transitive Alignments might face challenges when aligning MS/MS spectra of structurally related compounds with a significant difference in fragmentation mechanism. This may drastically alter MS/MS intensity, causing low aligned cosine scores, resulting in no putative connections for Transitive Alignments to build upon.

However, it is important to note that while a high density of single modifications can improve the performance of Transitive Alignments, it is not necessary for all single modifications to be present in the dataset for Transitive Alignments to be effective. The key-path concept of Transitive Alignments can tolerate missing intermediates as long as there is at least one path to form a connected component as a spanning tree. Further, we hypothesize that the completeness of data can be enhanced when analyzed from multiple samples within a dataset.

While we have demonstrated that the transitive alignment approach does improve the accuracy of MN construction, we note that there still exists a gap between what we can achieve and the upper bound true performance. This is hypothesized due to both limitations in structural similarity algorithms and how MS/MS fragmentation serves only as a proxy for structure. In Figure 2, we note the gap between the best performing practical network construction method (using Transitive Alignment + CAST) and the theoretically perfect network accuracy is less than 5% in low N20 value (overall MN component size is small) and for high N20 value (methods that try to form larger components) the practical maximum network accuracy score can only reach 50–70% of the ground truth network accuracy score.

It should also be noted that while Tanimoto similarity is the most frequently used way to compare molecular fingerprints, there are other metrics such as Dice similarities are occasionally used[14]. We also evaluated our benchmarking results using Dice similarity. As shown in SI Figure 5, the benchmarking results for the FDA Pt2 and NIH SPAC libraries reveal that the overall trend and networking performance of different network filtering methods remain consistent across both Tanimoto and Dice similarity metrics.

While we focus here on topology filtering to enhance the performance of modified cosine, we recognize the availability of other scoring metrics (MS2DeepScore included here) that aim to address the multiple modifications challenge. Specific examples including SIMILE[19] and Core Structure Search[8] were not evaluated here. However, looking forward, we envision that new topology filtering algorithms that are specific and complementary to a scoring metric will arise. Our analysis shows that the state-of-the-art techniques have not reached the upper bound of performance. We estimated that upper bound using the true chemical structures (Figure 2) and even with these advances introduced in this manuscript, there is a significant gap in performance. We hope that the benchmarking framework introduced here will provide a foundation to compare new combinations of similarity metrics and topology filtration algorithms to drive the community towards achieves the estimated optimal performance for MN construction.

Methods

Network Accuracy Metric

The metric we designed to measure accuracy is called the Network Accuracy Score. To compute this score, we first calculate the *Component Accuracy Score*. Suppose we have two nodes u and v, in a certain component, and they are connected by an edge, denoted as $edge_i$, resulting from the raw pairwise construction network topology. We define the $score\_edge_i$ as the structural similarity between molecules u and v that are connected by i. In this context, the *Component Accuracy Score* for a component is calculated as the sum of $score\_edge_i$ values divided by the total number of edges within the component—expressed mathematically as:

$$Component_{accuracy_{score}} = \frac{\sum_{i=1}^{m} score_{edge_i}}{|m|}, \qquad \#(1)$$

suppose we have m edges in the component. With the *Component Accuracy Score*, we calculate the Network Accuracy Score. Network Accuracy Score looks at the whole network, considering both the number of nodes in a component and how accurate that component is. This formulation is mathematically represented as:

$$Network_{Accuracy\_Score} = \frac{\sum_{k=1}^{n} Component_{accuracy_{score_k}} * num_{nodes_k}}{Total\_num_{nodes}} \qquad \#(2)$$

we have n number of components in the entire network.

Structure Similarity Calculation

To calculate the structure similarity score, we first convert the SMILES[20] or InChI[21] representations of molecules into RDKFingerprint, similar to Daylight fingerprints, using RDKit[22] (version 2023.09.2) with default settings: minimum path size of 1 bond, maximum path size of 7 bonds, fingerprint size of 2048 bits, number of bits set per hash of 2, minimum fingerprint size of 64 bits, and target on-bit density of 0.3. Then we use the Tanimoto[23] similarity metric to calculate the structure similarity for pairs of fingerprints.

Network Component Size Metric

To measure how big and comprehensive the constructed network is, we designed the Nth Number metric. It’s inspired by a widely utilized measurement in genomics called N50[24]. Drawing inspiration from the N50 concept, we adapted it to measure the size of components across an entire network. We sort the components by size from largest to smallest to calculate the Nth number. Then, add up nodes from the largest one until reach the N percentile of the entire network. The smallest number of nodes in the added component is the Nth number (i.e., the number of nodes in the last component that add up to reach the N percentile of the total number of nodes in the network). We’re not using the commonly used

50% threshold like in genomics because Molecular Networking has different characteristics. There are often a few significant parts and many smaller or single ones. We usually choose N20 because it captures the most important part of the features across most datasets.

Calculation of MS/MS Similarity

We calculated the modified cosine score in our experiments using the GNPS spectrum similarity method as the default setting[7]. The MS2 fragment ion tolerance is set to 0.5, the minimum matched peaks threshold is set to 3 and the maximum mass shift is set to 200.

Comparison Network Topology Algorithms

In this study, we focused on evaluating three fundamental methods: the baseline algorithm, the classic algorithm, and the CAST algorithm. We used our proposed metrics to assess their efficacy. Beyond examining existing approaches, we also introduced new algorithms to address identified challenges. Through this analysis, we aimed to provide a clear understanding of the strengths and limitations of each method, contributing to advancements in MN construction.

1. Baseline Algorithm

   The Baseline algorithm served as a foundation to eliminate potential biases and establish a reference point. This approach benchmarked the performance of the raw pairwise construction. The raw pairs construction was generated within the GNPS platform using the default settings: ms2_tolerance (0.5), min_cosine (0.7), and min_matched_peaks (3). For our benchmarking for the Baseline algorithm, we changed one main parameter, min_cosine, from 0.4 to 0.9 in increments of 0.1.

2. Classic Algorithm

   The GNPS Classic Method[6,7], operated as a Heuristic topology technique. This approach used the hyper-parameters of Top K and Max Component Size. This method consisted of two steps. The first step involved using the Top K nearest neighbors of each node, filtering out the low-scoring edges that exceed Top K. In the second step, the network was cut into components to meet the max component size criteria with the lowest scoring edges being removed first until the component was below the maximum component size. The default configuration for the GNPS Classic Method employed Top K set at 10 and Max Component Size set at 100. For our experiments, we adjusted Top K values from 1 to 40 in increments of 2 and iterate over the Max component parameters from 2 to 102 in increments of 5.

3. CAST Algorithm

   The Modified CAST Algorithm[13], performed pseudo clique identification via random optimization. The CAST algorithm started with the node which had the highest degree, then greedily grows the component from the nodes around it until it could no longer maintain the pseudo clique structure, or the average similarity (affinity) of the component fell below the set threshold. The primary parameter was the average similarity score threshold for components. Following this, the algorithm recursively processed all other nodes in the network until each node was assigned to its respective pseudo-clique. For the CAST algorithm, we swept through the threshold parameter range from 0.7 to 0.95 in increments of 0.01.

MS2DeepScore Retraining/Calculations

We retrained an MS2DeepScore model following the methods in the original MS2DeepScore publication[14]. Our approach involved adhering to the same preprocessing and datasets as described in the MS2DeepScore paper, except for excluding our benchmarking datasets (FDA Pt2, NIH SPAC, NIH NP, NIH NP Rd2 Positive, EMBL MCF, PSU-MSMLS — See Table 1) during the training stage. 500 structures were held out at random from the training set to use as validation data, with stereochemistry being ignored. This resulted in 62,664 training, 3,265 validation, and 43,585 test MS/MS spectra. Consistent with the publication, the MS2DeepScore model was trained using Jaccard similarities on RDKit Daylight Fingerprint and the Adam optimizer[25,26] to minimize the mean squared error (MSE) loss. A batch size of 32 and a learning rate of 0.001 were employed. Once retrained, we constructed the all-pairwise network using the MS2DeepScore. We filtered the MS2DeepScore network by sweeping a minimum similarity threshold from 0.4 to 0.99 in 0.01 increments.

Optimal Network Topology (structure similarity baseline)

The optimal network topology method is considered as the upper bound or ground truth for the construction of a MN on a certain dataset that we can achieve. It is built by calculating all pairwise actual structure similarity (see methods Structure Similarity Calculation) score instead of cosine score of the network and then set different filtering threshold parameter during benchmark from 0.4 to 0.9 in increments of 0.1 as we did for the baseline method.

Transitive Alignment Approach

To address the issue outlined in the introduction section, as illustrated in Figure 7, we draw inspiration from proteomics[27] and propose the Transitive Alignment approach. The process of Transitive Alignment approach is outlined in Figure 1.c. First, Transitive Alignment algorithm uses the single source shortest path (SSSP) algorithm[28] to determine all paths with the fewest hops between two MS/MS that are not directly connected in the network. In the next step, within the fewest hops paths find the one with highest sum of cosine similarity scores—a crucial path that we call the “key-path.” After finding the key path (e.g., A-D1-D2-B in Figure 1.c), we calculate the shifted peak sets of A-D1, D1-D2, and D2-B. We then leverage the connectivity transitivity of these three sets. For example, if peak i in A aligns with peak j in D1, and peak j in D1 aligns with peak k in D2, we can deduce that peak i in A can align with peak k in D2. This information is used to realign the peaks from A and B. This process captures the essence of our algorithm—using the intermediate nodes to readjust the relationship between two indirectly connected nodes, effectively dealing with challenges that arise from scenarios with multiple modifications.

CAST + Transitive Alignment Approach

In order to explore how Transitive Alignment Approach could benefit other filtering algorithms in MN construction, we proposed the CAST + Transitive Alignment Approach. The main process consists of three parts. First, we implemented the Transitive Alignment to reconstruct the MN based on the raw pairwise network. Then, we applied the CAST algorithm on this reconstructed MN to get the filtered MN. Finally, we conducted a modified greedy Maximum Spanning Tree (MST)[29] refinement to each component of the filtered MN to enhance the interpretability and clarity. For the CAST + Transitive Alignment approach, we swept through the threshold parameter, ranging from 0.7 to 0.95 in increments of 0.01, in the same manner as the CAST algorithm.

Benchmarking Datasets and Scoring Methods

Our investigation begins with the application of our proposed metrics across a diverse array of datasets, specifically 6 GNPS reference MS/MS libraries in Table 1.

We calculate Network Density[30] as the ratio of observed edges to the total potential edges within a network, offering a measure of interconnectedness. It’s mathematically defined as $2\,|E|/(|V|\,(|V|-1)$, where $|E|$ is the number of edges and $|V|$ represents the cardinality of vertices. This metric is computed based on the raw pairwise network topology.

Random Sampling Method

We initially sampled the network based on node count, applying different sample rates from 0.5 to 0.9 in increments of 0.2 to simulate various levels of network sparsity. Subsequently, we subjected these sampled networks to the same benchmarking protocol, allowing us to systematically analyze the impact of network sparsity on algorithmic performance while maintaining a controlled experimental setting. For each sample rate, we conducted the experiment 10 times and calculated the mean results.

Consistency measurement with ClassyFire

To evaluate the component sets of MN produced by filtering algorithms are chemically meaning, we conducted an approach by calculating the ratio of correctly classified component in the whole MN under the “superclass” level given by ClassyFire[17,18]. Whether a component is correctly classified is defined by if the purity of the component is above the setting threshold (we set to 0.7 in our experiment). Suppose we have a component contains n molecules in it, the most frequent annotation type is C by ClassyFire, the purity is calculated by number of the most frequent annotation type C counts k divided by the number of molecules in the component. It can be mathematically expressed as:

$$purity = \frac{k}{n} \qquad \#(3)$$

The correctly classified component count for i-th component is expressed as:

$$correctly\ \ classified\ \ component_i = \begin{cases} 0, & purity < 0.7 \\ 1, & purity \geq 0.7 \end{cases}$$

#(4)

So, the ratio of correctly classified component is the number of correctly classified component divided by the total number of components exist in the MN, which can be expressed as:

$$\text{ratio of correctly classified component} = \frac{\sum_{1}^{n} Correctly\ \ classified\ \ component_i}{\#\ \{components\}}$$

#(5)

Transitive Induced Networking

To create induced networks using the transitive alignment approach, we require a target node S that is present in the molecular network, max hops, and minimum transitive alignment score threshold from the user. From node S, induced networking employs the transitive alignment approach to realign all nodes not directly linked to it using the path derived from the SSSP algorithm with the maximum number of hops in the original raw pairwise edges and adds in transitive edges that are above the minimum transitive alignment score threshold. Subsequently, leveraging the network topology from the previous step, we induce the network starting from the node S. During this induction process, a Breadth First Search (BFS) node traversal from node S is carried out first within the specified maximum number of hops (which indicates the maximum number of potential structure modifications). The nodes visited during this traversal are denoted as the node set C. In the final step, we extract the subnetwork formed by the nodes that satisfy two requirements: their existence within node set C, and the transitive alignment scores between them and node S are above the set threshold.

Corteva Mass Spectrometry Experimental Methods

Corteva fungal strain DF978Z0035 was cultured in CF25ST for two weeks and extracted with EtOAC. Aliquot of the extract was dried down and redissolved in MeOH. Untargeted mass spectrometry data was acquired on an Thermo ID-X Tribrid mass spectrometry with Thermo Vanquish UPLC system (San Jose, CA). HPLC settings: Waters Cortecs column, 1.6 μm, 2.1 X 50 mm, solvent A = 0.1% formic acid in water, B = 0.1% formic acid in acetonitrile, flow rate 0.4 ml/min. Two uL of extract in MeOH was injected and eluted with 5% B for 1 min, followed by linear gradient to 100% B in 10 min. DDA settings: MS1 resolution 60000 (profile), MS2 resolution 30000 (centroid); scan range 150 – 2000 m/z, cycle time 0.6 second, dynamic exclusion times 1 and exclusion duration 2.5 second. Normalized collision was conducted in HCD (20, 40 and 60).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We would like to thank Nuno Bandeira for early fruitful discussions. We would also like to acknowledge Serge Fotso for the isolation and structure elucidation of porritoxin analogs.

XW is supported by a sponsored research agreement with Corteva Agriscience. MW is partially supported by NIH 5U24DK133658–02, UCR Startup Funds, and the U.S. Department of Energy Joint Genome Institute (https://ror.org/04xm1d337), a DOE Office of Science User Facility, is supported by the Office of Science of the U.S. Department of Energy operated under Contract No. DE-AC02–05CH11231. MS is partially supported by the U.S. Department of Energy Joint Genome Institute (https://ror.org/04xm1d337), a DOE Office of Science User Facility, is supported by the Office of Science of the U.S. Department of Energy operated under Contract No. DE-AC02–05CH11231. ATA was supported by funds from the Boettcher Foundation’s Webb-Waring Biomedical Research Program and DU Startup Funds. VVP was supported by NIH R35GM128690 and the ALSAM Foundation.

Data Availability

Zenodo : https://zenodo.org/records/10724765

GNPS2 Task for Corteva Analysis Links

Classic Network Task: https://gnps2.org/status?task=3232258d36124077b4dc3e017a8103d6

Induced Network Task: https://gnps2.org/status?task=86dad9a8ac764417a850e3fb8dbe3d18

Bibliography

(1). Lima NM; Dos Santos GF; da Silva Lima G; Vaz BG Advances in Mass Spectrometry-Metabolomics Based Approaches. Adv. Exp. Med. Biol 2023, 1439, 101–122. 10.1007/978-3-031-41741-2_5. [PubMed: 37843807]

(2). Neumann S; Böcker S Computational Mass Spectrometry for Metabolomics: Identification of Metabolites and Small Molecules. Anal. Bioanal. Chem 2010, 398 (7–8), 2779–2788. 10.1007/s00216-010-4142-5. [PubMed: 20936272]

(3). Patti GJ; Yanes O; Siuzdak G Metabolomics: The Apogee of the Omics Trilogy. Nat. Rev. Mol. Cell Biol 2012, 13 (4), 263–269. 10.1038/nrm3314. [PubMed: 22436749]

(4). Huan T; Palermo A; Ivanisevic J; Rinehart D; Edler D; Phommavongsay T; Benton HP; Guijas C; Domingo-Almenara X; Warth B; Siuzdak G Autonomous Multimodal Metabolomics Data Integration for Comprehensive Pathway Analysis and Systems Biology. Anal. Chem 2018, 90 (14), 8396–8403. 10.1021/acs.analchem.8b00875. [PubMed: 29893550]

(5). Guijas C; Montenegro-Burke JR; Domingo-Almenara X; Palermo A; Warth B; Hermann G; Koellensperger G; Huan T; Uritboonthai W; Aisporna AE; Wolan DW; Spilker ME; Benton HP; Siuzdak G METLIN: A Technology Platform for Identifying Knowns and Unknowns. Anal. Chem 2018, 90 (5), 3156–3164. 10.1021/acs.analchem.7b04424. [PubMed: 29381867]

(6). Watrous J; Roach P; Alexandrov T; Heath BS; Yang JY; Kersten RD; van der Voort M; Pogliano K; Gross H; Raaijmakers JM; Moore BS; Laskin J; Bandeira N; Dorrestein PC Mass Spectral Molecular Networking of Living Microbial Colonies. Proc. Natl. Acad. Sci. U. S. A 2012, 109 (26), E1743–1752. 10.1073/pnas.1203689109. [PubMed: 22586093]

(7). Wang M; Carver JJ; Phelan VV; Sanchez LM; Garg N; Peng Y; Nguyen DD; Watrous J; Kapono CA; Luzzatto-Knaan T; Porto C; Bouslimani A; Melnik AV; Meehan MJ; Liu W-T; Crüsemann M; Boudreau PD; Esquenazi E; Sandoval-Calderón M; Kersten RD; Pace LA; Quinn RA; Duncan KR; Hsu C-C; Floros DJ; Gavilan RG; Kleigrewe K; Northen T; Dutton RJ; Parrot D; Carlson EE; Aigle B; Michelsen CF; Jelsbak L; Sohlenkamp C; Pevzner P; Edlund A; McLean J; Piel J; Murphy BT; Gerwick L; Liaw C-C; Yang Y-L; Humpf H-U; Maansson M; Keyzers RA; Sims AC; Johnson AR; Sidebottom AM; Sedio BE; Klitgaard A; Larson CB; P, C. A. B.; Torres-Mendoza D; Gonzalez DJ; Silva DB; Marques LM; Demarque DP; Pociute E; O’Neill EC; Briand E; Helfrich EJN; Granatosky EA; Glukhov E; Ryffel F; Houson H; Mohimani H; Kharbush JJ; Zeng Y; Vorholt JA; Kurita KL; Charusanti P; McPhail KL; Nielsen KF; Vuong

L; Elfeki M; Traxler MF; Engene N; Koyama N; Vining OB; Baric R; Silva RR; Mascuch SJ; Tomasi S; Jenkins S; Macherla V; Hoffman T; Agarwal V; Williams PG; Dai J; Neupane R; Gurr J; Rodríguez AMC; Lamsa A; Zhang C; Dorrestein K; Duggan BM; Almaliti J; Allard P-M; Phapale P; Nothias L-F; Alexandrov T; Litaudon M; Wolfender J-L; Kyle JE; Metz TO; Peryea T; Nguyen D-T; VanLeer D; Shinn P; Jadhav A; Müller R; Waters KM; Shi W; Liu X; Zhang L; Knight R; Jensen PR; Palsson BO; Pogliano K; Linington RG; Gutiérrez M; Lopes NP; Gerwick WH; Moore BS; Dorrestein PC; Bandeira N Sharing and Community Curation of Mass Spectrometry Data with Global Natural Products Social Molecular Networking. Nat. Biotechnol 2016, 34 (8), 828–837. 10.1038/nbt.3597. [PubMed: 27504778]

(8). Xing S; Hu Y; Yin Z; Liu M; Tang X; Fang M; Huan T Retrieving and Utilizing Hypothetical Neutral Losses from Tandem Mass Spectra for Spectral Similarity Analysis and Unknown Metabolite Annotation. Anal. Chem 2020, 92 (21), 14476–14483. 10.1021/acs.analchem.0c02521. [PubMed: 33076659]

(9). Shen X; Wang R; Xiong X; Yin Y; Cai Y; Ma Z; Liu N; Zhu Z-J Metabolic Reaction Network-Based Recursive Metabolite Annotation for Untargeted Metabolomics. Nat. Commun 2019, 10 (1), 1516. 10.1038/s41467-019-09550-x. [PubMed: 30944337]

(10). Moorthy AS; Wallace WE; Kearsley AJ; Tchekhovskoi DV; Stein SE Combining Fragment-Ion and Neutral-Loss Matching during Mass Spectral Library Searching: A New General Purpose Algorithm Applicable to Illicit Drug Identification. Anal. Chem 2017, 89 (24), 13261–13268. 10.1021/acs.analchem.7b03320. [PubMed: 29156120]

(11). Burke MC; Mirokhin YA; Tchekhovskoi DV; Markey SP; Heidbrink Thompson J; Larkin C; Stein SE The Hybrid Search: A Mass Spectral Library Search Method for Discovery of Modifications in Proteomics. J. Proteome Res 2017, 16 (5), 1924–1935. 10.1021/acs.jproteome.6b00988. [PubMed: 28367633]

(12). Bittremieux W; Schmid R; Huber F; Van Der Hooft JJJ; Wang M; Dorrestein PC Comparison of Cosine, Modified Cosine, and Neutral Loss Based Spectrum Alignment For Discovery of Structurally Related Molecules. J. Am. Soc. Mass Spectrom 2022, 33 (9), 1733–1744. 10.1021/jasms.2c00153. [PubMed: 35960544]

(13). Ben-Dor A; Shamir R; Yakhini Z Clustering Gene Expression Patterns. J. Comput. Biol. J. Comput. Mol. Cell Biol 1999, 6 (3–4), 281–297. 10.1089/106652799318274.

(14). Huber F; van der Burg S; van der Hooft JJJ; Ridder L MS2DeepScore: A Novel Deep Learning Similarity Measure to Compare Tandem Mass Spectra. J. Cheminformatics 2021, 13 (1), 84. 10.1186/s13321-021-00558-4.

(15). Guthals A; Watrous JD; Dorrestein PC; Bandeira N The Spectral Networks Paradigm in High Throughput Mass Spectrometry. Mol. Biosyst 2012, 8 (10), 2535–2544. 10.1039/c2mb25085c. [PubMed: 22610447]

(16). The M; Käll L MaRaCluster: A Fragment Rarity Metric for Clustering Fragment Spectra in Shotgun Proteomics. J. Proteome Res 2016, 15 (3), 713–720. 10.1021/acs.jproteome.5b00749. [PubMed: 26653874]

(17). Kim HW; Wang M; Leber CA; Nothias L-F; Reher R; Kang KB; van der Hooft JJJ; Dorrestein PC; Gerwick WH; Cottrell GW NPClassifier: A Deep Neural Network-Based Structural Classification Tool for Natural Products. J. Nat. Prod 2021, 84 (11), 2795–2807. 10.1021/acs.jnatprod.1c00399. [PubMed: 34662515]

(18). Djoumbou Feunang Y; Eisner R; Knox C; Chepelev L; Hastings J; Owen G; Fahy E; Steinbeck C; Subramanian S; Bolton E; Greiner R; Wishart DS ClassyFire: Automated Chemical Classification with a Comprehensive, Computable Taxonomy. J. Cheminformatics 2016, 8, 61. 10.1186/s13321-016-0174-y.

(19). Treen DGC; Wang M; Xing S; Louie KB; Huan T; Dorrestein PC; Northen TR; Bowen BP SIMILE Enables Alignment of Tandem Mass Spectra with Statistical Significance. Nat. Commun 2022, 13 (1), 2510. 10.1038/s41467-022-30118-9. [PubMed: 35523965]

(20). Weininger D SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. J. Chem. Inf. Comput. Sci 1988, 28 (1), 31–36. 10.1021/ci00057a005.

(21). Heller SR; McNaught A; Pletnev I; Stein S; Tchekhovskoi D InChI, the IUPAC International Chemical Identifier. J. Cheminformatics 2015, 7 (1), 23. 10.1186/s13321-015-0068-4.

(22). Landrum Greg; Tosco Paolo; Kelley Brian; Ric; Cosgrove David; sriniker; gedeck; Vianello Riccardo; Schneider Nadine; Kawashima Eisuke; Jones Gareth; Dan N; Dalke Andrew; Cole Brian; Swain Matt; Turk Samo; Savelyev Alexander; Vauche Alain; Wójcikowski Maciej; Take chiru; Probst Daniel; Ujihara Kazuya; Scalfani Vincent F.; godin guillaume; Walker Rachel; Lehtivarjo Juuso; Pahl Axel; Berenger Francois; jasondbiggs; strets123. Rdkit/Rdkit: 2023_09_2 (Q3 2023) Release, 2023. 10.5281/ZENODO.10099869.

(23). Willett P Similarity-Based Virtual Screening Using 2D Fingerprints. Drug Discov. Today 2006, 11 (23–24), 1046–1053. 10.1016/j.drudis.2006.10.005. [PubMed: 17129822]

(24). Lander ES; Linton LM; Birren B; Nusbaum C; Zody MC; Baldwin J; Devon K; Dewar K; Doyle M; FitzHugh W; Funke R; Gage D; Harris K; Heaford A; Howland J; Kann L; Lehoczky J; LeVine R; McEwan P; McKernan K; Meldrim J; Mesirov JP; Miranda C; Morris W; Naylor J; Raymond C; Rosetti M; Santos R; Sheridan A; Sougnez C; Stange-Thomann N; Stojanovic N; Subramanian A; Wyman D; The Sanger Centre:; Rogers J; Sulston J; Ainscough R; Beck S; Bentley D; Burton J; Clee C; Carter N; Coulson A; Deadman R; Deloukas P; Dunham A; Dunham I; Durbin R; French L; Grafham D; Gregory S; Hubbard T; Humphray S; Hunt A; Jones M; Lloyd C; McMurray A; Matthews L; Mercer S; Milne S; Mullikin JC; Mungall A; Plumb R; Ross M; Shownkeen R; Sims S; Washington University Genome Sequencing Center; Waterston RH; Wilson RK; Hillier LW; McPherson JD; Marra MA; Mardis ER; Fulton LA; Chinwalla AT; Pepin KH; Gish WR; Chissoe SL; Wendl MC; Delehaunty KD; Miner TL; Delehaunty A; Kramer JB; Cook LL; Fulton RS; Johnson DL; Minx PJ; Clifton SW; US DOE Joint Genome Institute:; Hawkins T; Branscomb E; Predki P; Richardson P; Wenning S; Slezak T; Doggett N; Cheng J-F; Olsen A; Lucas S; Elkin C; Uberbacher E; Frazier M; Baylor College of Medicine Human Genome Sequencing Center:; Gibbs RA; Muzny DM; Scherer SE; Bouck JB; Sodergren EJ; Worley KC; Rives CM; Gorrell JH; Metzker ML; Naylor SL; Kucherlapati RS; Nelson DL; Weinstock GM; RIKEN Genomic Sciences Center:; Sakaki Y; Fujiyama A; Hattori M; Yada T; Toyoda A; Itoh T; Kawagoe C; Watanabe H; Totoki Y; Taylor T; Genoscope and CNRS UMR-8030:; Weissenbach J; Heilig R; Saurin W; Artiguenave F; Brottier P; Bruls T; Pelletier E; Robert C; Wincker P; Department of Genome Analysis, Institute of Molecular Biotechnology:; Rosenthal A; Platzer M; Nyakatura G; Taudien S; Rump A; GTC Sequencing Center:; Smith DR; Doucette-Stamm L; Rubenfield M; Weinstock K; Lee HM; Dubois J; Beijing Genomics Institute/ Human Genome Center:; Yang H; Yu J; Wang J; Huang G; Gu J Multimegabase Sequencing Center, The Institute for Systems Biology:; Hood L; Rowen L; Madan A; Qin S; Stanford Genome Technology Center:; Davis RW; Federspiel NA; Abola AP; Proctor MJ; University of Oklahoma’s Advanced Center for Genome Technology:; Roe BA; Chen F; Pan H; Max Planck Institute for Molecular Genetics:; Ramser J; Lehrach H; Reinhardt R; Cold Spring Harbor Laboratory, Lita Annenberg Hazen Genome Center:; McCombie WR; De La Bastide M; Dedhia N; GBF—German Research Centre for Biotechnology:; Blöcker H; Hornischer K; Nordsiek G; *Genome Analysis Group (listed in alphabetical order, also includes individuals listed under other headings):; Agarwala R; Aravind L; Bailey JA; Bateman A; Batzoglou S; Birney E; Bork P; Brown DG; Burge CB; Cerutti L; Chen H-C; Church D; Clamp M; Copley RR; Doerks T; Eddy SR; Eichler EE; Furey TS; Galagan J; Gilbert JGR; Harmon C; Hayashizaki Y; Haussler D; Hermjakob H; Hokamp K; Jang W; Johnson LS; Jones TA; Kasif S; Kaspryzk A; Kennedy S; Kent WJ; Kitts P; Koonin EV; Korf I; Kulp D; Lancet D; Lowe TM; McLysaght A; Mikkelsen T; Moran JV; Mulder N; Pollara VJ; Ponting CP; Schuler G; Schultz J; Slater G; Smit AFA; Stupka E; Szustakowki J; Thierry-Mieg D; Thierry-Mieg J; Wagner L; Wallis J; Wheeler R; Williams A; Wolf YI; Wolfe KH; Yang S-P; Yeh R-F; Scientific management: National Human Genome Research Institute, US National Institutes of Health:; Collins F; Guyer MS; Peterson J; Felsenfeld A; Wetterstrand KA; Stanford Human Genome Center:; Myers RM; Schmutz J; Dickson M; Grimwood J; Cox DR; University of Washington Genome Center:; Olson MV; Kaul R; Raymond C; Department of Molecular Biology, Keio University School of Medicine:; Shimizu N; Kawasaki K; Minoshima S; University of Texas Southwestern Medical Center at Dallas:; Evans GA; Athanasiou M; Schultz R; Office of Science, US Department of Energy:; Patrinos A; The Wellcome Trust:; Morgan MJ Initial Sequencing and Analysis of the Human Genome. Nature 2001, 409 (6822), 860–921. 10.1038/35057062. [PubMed: 11237011]

(25). Goodfellow I; Bengio Y; Courville A Deep Learning; Adaptive computation and machine learning; The MIT Press: Cambridge, Massachusetts, 2016.

(26). Kingma DP; Ba J Adam: A Method for Stochastic Optimization. arXiv January 29, 2017. http://arxiv.org/abs/1412.6980 (accessed 2024-02-02).

(27). Bandeira N; Tsur D; Frank A; Pevzner PA Protein Identification by Spectral Networks Analysis. Proc. Natl. Acad. Sci. U. S. A 2007, 104 (15), 6140–6145. 10.1073/pnas.0701130104. [PubMed: 17404225]

(28). Dijkstra EW A Note on Two Problems in Connexion with Graphs. Numer. Math 1959, 1 (1), 269–271. 10.1007/BF01386390.

(29). Kruskal JB On the Shortest Spanning Subtree of a Graph and the Traveling Salesman Problem. Proc. Am. Math. Soc 1956, 7 (1), 48–50. 10.1090/S0002-9939-1956-0078686-7.

(30). Coleman TF; Moré JJ Estimation of Sparse Jacobian Matrices and Graph Coloring Blems. SIAM J. Numer. Anal 1983, 20 (1), 187–209. 10.1137/0720013.

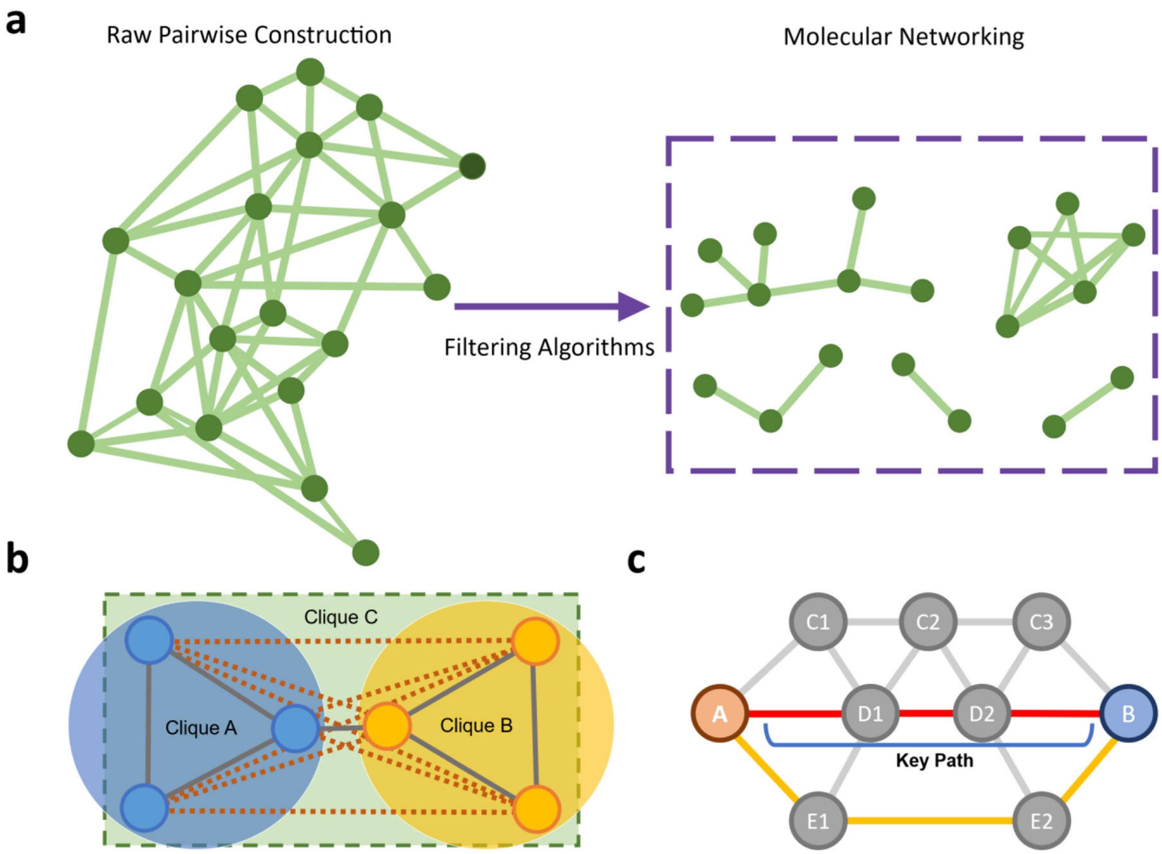


Figure 1 –. Construction of Molecular Networking and Transitive Alignment Overview

a) Constructing Molecular Networks involves two steps: first, computing pairwise similarity scores to form the foundational network, and then applying a filtering algorithm to enhance clarity and meaningful connections. b) In this example, the CAST algorithm can only identify small Cliques—Clique A and B—in the original raw pairwise MN. The missing edges (dashed red edges) are caused by multiple modifications which can cause clique finding algorithms to produce very small cliques. By adding back edges, the initial Clique A and B can reform into a clique double in size, forming Clique C in the green dashed box. c) The schematic of the Transitive Alignments. First, to align MS/MS spectra A and B, we find all single source shortest paths between A and B (called the key path). If there are many key paths, we choose the one with maximum sum of cosine score along the path. Next, we use each intermediate node along the key path to realign the MS/MS peaks of A and B.

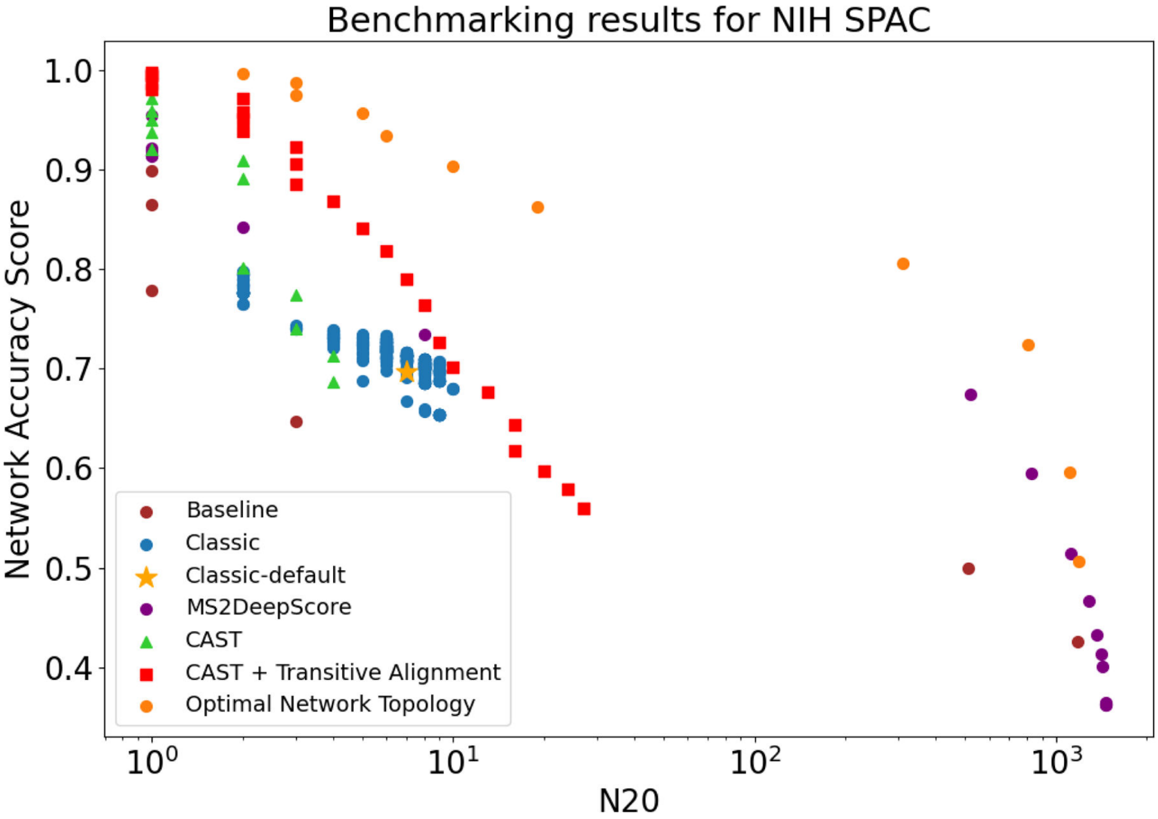


Figure 2 –.

Benchmarking results for topology of Pharmacologically Active Compounds in the NIH Small Molecule Repository data. The x-axis signifies the N20 number, where a larger value is more favorable, while the y-axis represents the network accuracy score—higher values are desirable. Hence, the ideal trend for improvement and optimization is characterized by a curve that trends toward the upper right corner. Each scatter point for the baseline denotes outcomes obtained under a specific min_cosine parameter setting. Similarly, each scatter point for the GNPS Classic method reflects results from varying combinations of top k and max component parameters. For CAST and CAST + Transitive Alignment, each scatter point corresponds to a certain threshold parameter setting (See Methods Comparison Network Topology Algorithms and CAST + Transitive Alignment Approach). We observe that all methods surpass the baseline in this dataset. The Classic method shows modest gains, while another pseudo clique identification algorithm, the CAST algorithm, excels at lower N20 values but lags the Classic method at higher N20 values. Notably, the CAST + Transitive Alignment method demonstrates the most substantial improvement compared to other methods. Benchmarking results for all the datasets see in SI Figure 1.

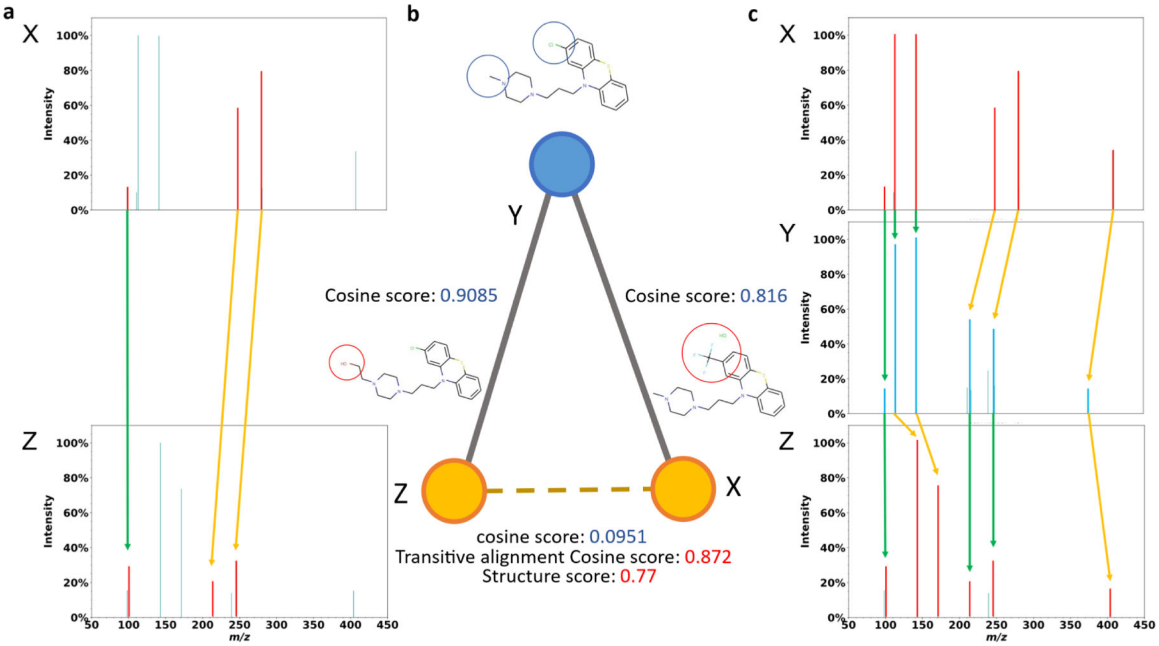


Figure 3 –. The effectiveness of the Transitive Alignment method.

a) The spectra and peak alignment of molecules X and Z using shifted peak alignment. b) A subgraph of molecules X, Y and Z extracted from the MN. The molecular structures of X, Y and Z are structurally similar, yet X and Z have an unexpectedly low modified cosine score of 0.0951, which does not align with their observed structural similarity. c) The MS/MS peak alignment of molecules X, Y and Z using Transitive Alignment approach. The shifted peak alignment method only aligns 3 peaks from X to Z in subgraph a, while the Transitive Alignment approach can align all top 6 (sorted by intensity) peaks from X to Z (c).

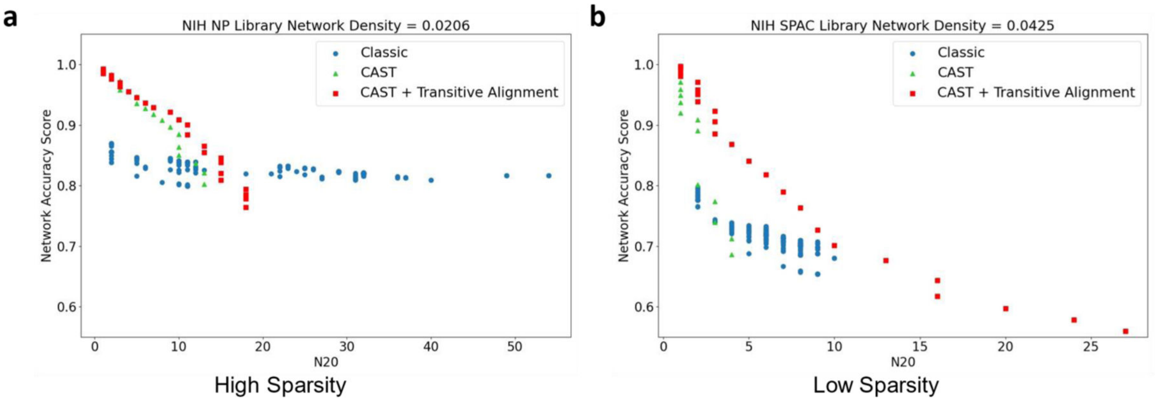


Figure 4 –.

Benchmarking results for Classic, CAST, CAST + alignment on different sparsity datasets. a) Results from NIH NP library, the network density for this dataset is 0.0206. b) Results from NIH SPAC library, the network density of this datasets is 0.0425. We can observe that in a low sparsity network, the improvement from applying Transitive alignment to CAST is larger than that in a high sparsity network.

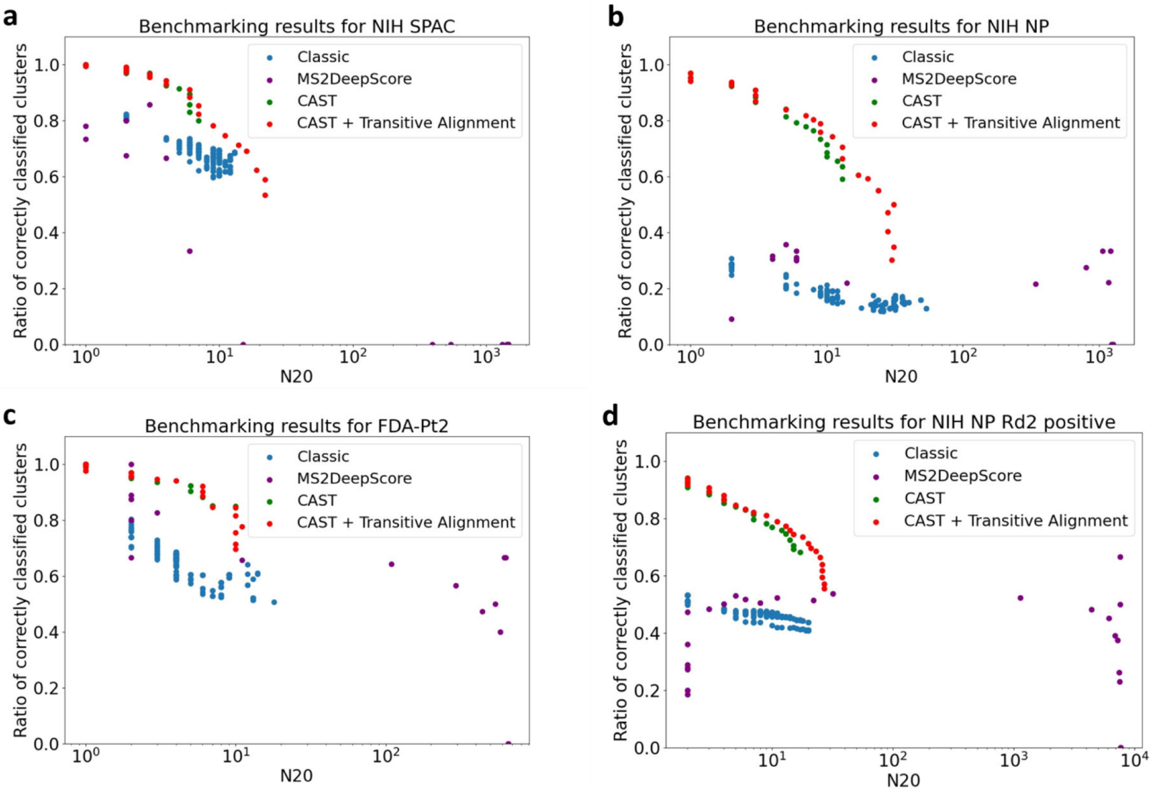


Figure 5 —.

Ratio of correctly classified clusters with N20 for different MN construction algorithms. A cluster’s classification correctness is determined by its purity being over 0.7. This is shown for (a) NIH SPAC, (b) NIH NP, (c) FDA-Pt2, and (d) NIH NP Rd2 positive, comparing the performance of Classic, MS2DeepScore, CAST, and CAST + Transitive Alignment methods. The x-axis represents N20, and the y-axis represents the ratio of correctly classified clusters. Across all datasets, the CAST method generally achieves the highest performance for small N20 values but declines as N20 increases.

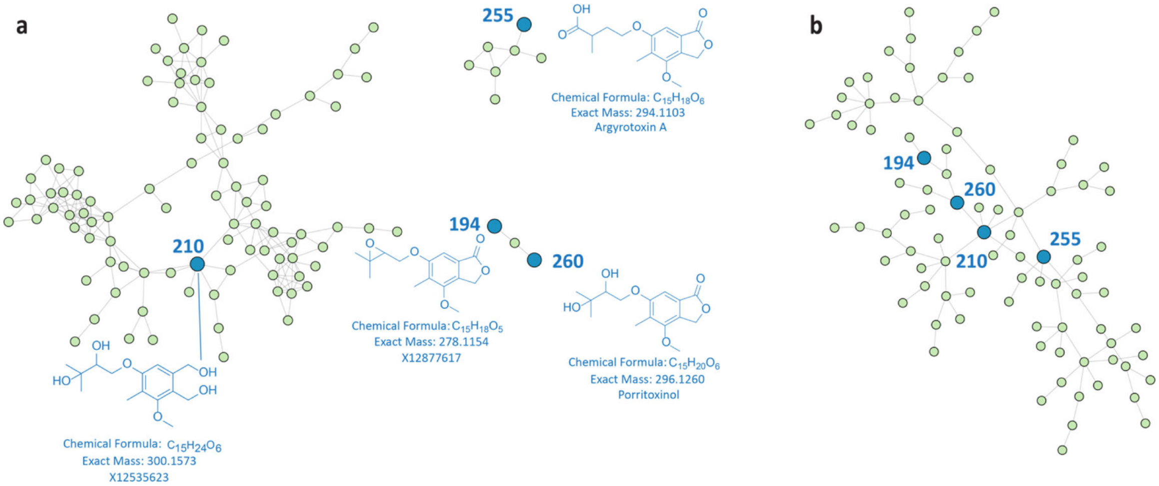


Figure 6 –.

Demonstration of transitive alignment creating more complete molecular networks of known similar molecules. Here we show an example of **a fungal strain Alternaria sp** (**DF978Z0035)** producing **porritoxins** class of molecules. Originally, four known members of this family of molecules were present in 3 components (a). With transitive alignments, all four are grouped into the same molecular family (b).

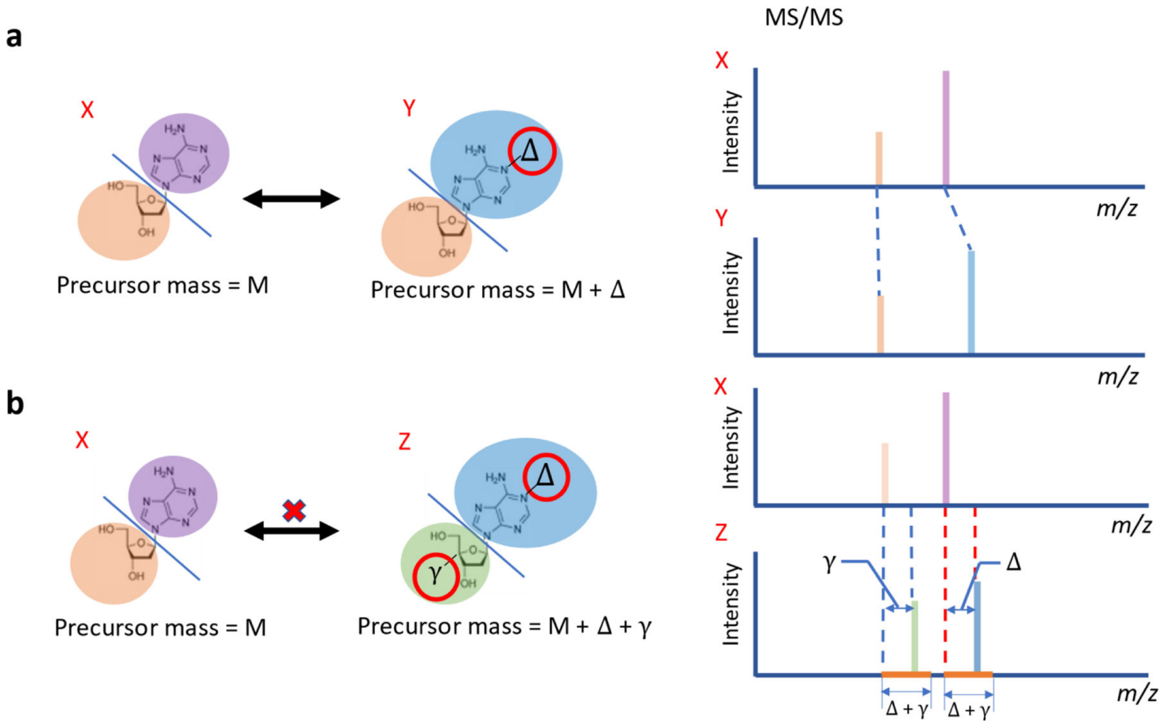


Figure 7 –.

Shifted alignment for single and multiple modifications. a) Molecules X and Y, alongside their respective MS/MS spectra. Molecule Y has one modification, Δ, compared to X. Suppose the precursor mass of A is denoted as M, and for molecule B, it becomes M + Δ. If we do not align the shifted peaks of the MS/MS data to calculate cosine scores of X and Y, it can yield low results because of mismatched fragmentation peak masses. However, the single modification should yield a high spectrum similarity score, accurately reflecting structural similarity. To overcome this hurdle, the common way is to aligns peaks in the MS/MS spectrum of molecule Y with those in molecule X if they are shifted within or near the precursor mass difference Δ. While shift alignment adeptly addresses single modifications, the complexity magnifies when dealing with multiple modifications. b) Molecules X and Z have a double modification with distinct masses, Δ and γ on different sites. This resulted in a total mass difference (Δ + γ) that complicates straightforward alignment, leading to inconsistencies in transitivity.

Table 1 –

Features of the 6 datasets we benchmarked, which were selected to cover a wide range of sizes and network densities. The full name and description of each dataset are in SI Table 1.

Abbreviation Library name	# of nodes	# of edges	Network Density	Instrument
FDA Pt2	513	10921	0.0831	qTof
NIH SPAC	1201	30669	0.0425	qTof
NIH NP	1093	12342	0.0206	qTof
NIH NP Rd2 Positive	7170	488865	0.0190	qTof
EMBL MCF	548	9915	0.0661	Orbitrap
PSU-MSMLS	476	1501	0.0132	qTof