Published in final edited form as:

Mol Omics. ; 17(5): 665–676. doi:10.1039/d1mo00005e.

ENHANCED DETECTION AND ANNOTATION OF SMALL MOLECULES IN METABOLOMICS USING MOLECULAR NETWORK-ORIENTED PARAMETER OPTIMIZATION

Rui Xu[1], **Jisun Lee**[1], **Li Chen**[1], **Jiangjiang Zhu**[1,2,#]

[1]Human Nutrition Program, The Ohio State University, Columbus, Ohio 43210

[2]James Comprehensive Cancer Center, The Ohio State University, Columbus, Ohio 43210

Abstract

Metabolomics, especially large-scale untargeted metabolomics, generate massive amounts of data on a regular basis, which often need to be filtered, screened, analyzed and annotated via a variety of approaches. Data-dependent acquisition (DDA) mode including inclusion and exclusion rules for tandem mass spectrometry (MS) is routinely used to perform such analyses. While parameters of data acquisition are important in these processes, there is a lack of systematic studies of these parameters that can be used in data collection to generate metabolic features for molecular network (MN) analysis on the Global Natural Product Social Molecular Networking platform (GNPS). To explore the key parameters that impacting the formation and quality of MNs, several data acquisition parameters for metabolomic studies were proposed in this study. The influences of MS^1 resolution, normalized collision energy (NCE), intensity threshold, exclusion time to GNPS analyses were demonstrated. Moreover, an optimization workflow dedicated to Thermo Scientific QE Hybrid Orbitrap instruments is described, and a comparison of phytochemical contents from two forms of black raspberry extracts were performed based on the GNPS MN results. Overall, we expect this study to provide additional thoughts on developing natural product analysis workflow using GNPS network, and shed some lights to future analyses that utilizing similar instrumental setups.

Keywords

Metabolomics; molecular network; parameter optimization; mass spectrometry

Introduction

Mass spectrometry (MS) is a powerful technique for metabolomics research to explore small molecular compounds and achieve extensive metabolite/natural product analysis [1]. In untargeted MS-based metabolomics, it remains a challenge to obtain high confidence in the identification/annotation of thousands of metabolites from the massive LC-MS /MS raw

[#]**To whom correspondence should be addressed to:** Jiangjiang (Chris) Zhu, Ph.D., Comprehensive Cancer Center and Department of Human Sciences, The Ohio State University, 400 W 12th Ave, Columbus, OH 43210 Tel: 614-685-2226, zhu.2484@osu.edu.

Disclosures: The authors have no conflict of interest to report data in the absence of prior knowledge [2]. In data-dependent acquisition (DDA), precursor signals in MS^1 spectra are interrogated and selected for fragmentation based on relative MS^1 signal intensity, giving rise to the MS^2 spectrum [3]. Then MS^2 spectra are matched to online database or in-home standards for identification. During the process, the extend of data acquisition determines the comprehensiveness of metabolic profile, and the identification precision determines the accuracy of the subsequent biological analysis. Increasing the number and accuracy of identification is a long-term goal for metabolomics [4]. There are many parameter settings in a DDA acquisition that affect the quantity and quality of MS^2 spectrum. For example, one study showed that meaningful MS^2 spectra were collected under optimized DDA conditions at both high and low mass resolutions by manipulating collision-induced dissociations [5]. Another study optimized the intensity threshold, fragmented peaks, and exclusion after N scans, which suggested that these parameters affected the MS^2 spectra quality under molecular network (MN) evaluation [6]. In addition, some specific libraries, such as the polyphenol database that are of interest in this study, are incomplete, which also influence the identification result. For decades, there is no open access database containing the structure of all natural products produced by plants, which hinders the development of compound identification [7].

Therefore, inferring unknown compounds from known compounds is considered as an effective way to improve identification in metabolomics analyses. Global Natural Product Social (GNPS), the first full open access resource of this type, provides a central repository for the formation of known natural products and a tool for MN generation [8]. MN is a visualized computing strategy that provides an intuitive view of all detected molecules, ions and the chemical relationships between these molecules and ions, playing an important role in large-scale compound identification and discovery of new compounds [9]. As similar MS^2 spectra can be generated from compounds with similar structures, molecules of the same class of compounds tend to converge into clusters of nodes during MN analysis, which is conducive to the qualitative analysis of unknown compounds [9]. Because of the intersectional connections between MNs, large-scale MS^2 spectrogram analysis is also possible. For example, one study combined the LipidXplorer glycoalkaloids list and GNPS analysis was used in Cytoscape to label nodes in the molecular network, triggered the structure elucidation of closely related nodes leading to the identification of 30 compounds using the LipidXplorer output and four purified and structure elucidated compounds[10]. Another study focused on identification of a subset of polyphenols in Casearia using GNPS, and for the first time identified eight compounds in the family of flavonoid-3-O-glycosides [11]. In addition, GNPS and molecular networks have also been shown to show cross-links between the chemistry of seemingly unrelated biological systems and used for drug discovery, drug metabolism and precision medicine [12]. Despite these progresses, it is also well-recognized that MN results can be influenced by MS^2 data quality as the MN is organized by spectral similarities [13], therefore, integrated MS parameter optimization and its consequential MN analysis should be carefully conducted to enable reliable MS^2 data generation and efficient MN construction.

In this study, we aim to optimize several important parameters on the Thermo QE hybrid MS platform that can potentially increase both quantity and quality of MS^2 spectra, so that enhanced natural product identification can be achieved. The parameters optimized in this study include the resolution of MS^1 resolution, normalized collision energy (NCE), intensity threshold, and exclusion time. These parameters of DDA were optimized based on GNPS MN results to generate better MS^2 spectra for annotation. The MNs formed by the collected MS^2 spectra were evaluated to determine the polyphenol compounds and derivatives of black raspberry (BRB) extracts in both liquid and powder forms from the same vendor. By comparing the optimization results of different data collection parameters, the importance of MS parameter optimizations for natural product analysis were also demonstrated.

Material and Methods

BRB Extract sample preparation

Both BerriHealth Premium Alcohol Free Black Raspberry Extract (BRB liquid) and Freeze-Dried Black Raspberry Powder (BRB powder) were purchased from Berrihealth company (Corvallis, OR, US) and used in this study. One hundred and fifty mg of liquid BRB extract or one hundred mg of powder BRB extract was dissolved in 1 ml solution (contains 20% HPLC water and 80% methanol). The mix were sonicated for 30 minutes at room temperature and then centrifuged for 10 minutes at 100g, 4°C. The supernatant was collected and filtered through 0.2 μm PTFE filter. The filtrate was immediately transferred to 1.5 ml LC-MS vials for detection.

Parameter setting on HPLC-Q-Exactive Mass Spectrometer system

LC-MS analysis was performed for prepared serum samples on an HPLC-Q-Exactive Mass Spectrometer system with XTERRA RP 18 Column (3.5 μm, 3.9 mm X 100 mm; Waters Corporation, Milford, MA) with gradient mobile phases for 12 minutes at a flow rate of 0.9 ml/min. Mobile phase A consisted of water and acetonitrile at a ratio of 9:1 (containing 5mM ammonium acetate and 0.1% acetic acid), mobile phase B consisted of water and acetonitrile at a ratio of 1:9 (containing 5mM ammonium acetate and 0.1% acetic acid). The injection gradients were as follows: mobile phase A was 99% at 0 minute, next decreased to 1% in 8 minutes, then increased to 99% in 2 minutes, finally kept 99% to 12 minutes. BRB extract samples were tested in both positive and negative electrospray ionization modes. In this experiment, the data acquisition mode was full scan+ $DDMS^2$, and high quality MS^1 and MS^2 data were obtained. The specific parameter setting was as below and shown in Table S1 in detail: MS^1 resolution (17,500, 35,000, 70,000,140,000), NCE (10, 40, 70 collision energy), intensity threshold (10E4, 10E5,10E6), exclusion time (10 s, 30 s, 60 s). A total of 432 injections were performed by applying the combination of these parameters in this study.

Molecular network analysis on GNPS

Proteo Wizard was applied to convert raw data to mzXL format prior to GNPS analysis. Converted data were then uploaded to GNPS [8] file transfer protocol (FTP, Host: ccms-ftp01.ucsd.edu) through an FTP client named Forklift. Then, on the MN job submission webpage of GNPS, the uploaded data were divided into groups by gradient according to the parameters explored. For each sample under either positive or negative mode, the converted data were submitted in four grouping methods (corresponding to four parameters). For the comparison of different extracts, the converted data of liquid and powder BRB extracts under the optimal parameter set, respectively, were submitted in 1 job with 2 groups. In the MN, the similarity between MS^2 spectra were calculated and represented by cosine score (range 0-1; the higher the score, the higher the spectral similarity). Any spectra above the specified threshold were then connected to form a visualized MN graph. In each polarity, when one parameter setting was discussed, data from all other 3 parameters were aligned and averaged for generalized comparison according to default GNPS setting (minimum matched fragment ions: 6; minimum pairs cosine score: 0.7) [14].

Data analysis and visualization

After peak extraction, feature identification and MN generation in GNPS, MN was exported to Cytoscape for further observation, annotation and analysis. In a MN, each circle represents a compound, and each edge represents a connection between two compounds (related by structure similarity identified via MS^2 spectra). Several networking statistics, including total nodes, identified nodes, nodes in network, self-loop nodes, pairs/edges, networks were collected directly. Networks (NW) is defined as total numbers of independent network clusters containing nodes >2. MN graphs were then downloaded for Cytoscape analysis [15] to calculate cosine score and cluster coefficient. Cosine score is the most widely used measures of spectral similarity [16]. Cluster coefficients (CC) is defined as the average measurement for “cliquishness” of the neighborhood of node [17]. These statistics were then summarized and used to generate eight-dimensional radar maps via MathWorks MATLAB (Natick, MA, USA). Cytoscape were also used to highlight nodes in different colors according to different parameters, samples, intensity levels or identification levels in MN diagrams. The mass spectra data has been deposited to MassIVE database (https://massive.ucsd.edu/ProteoSAFe/static/massive.jsp) with access # MSV000087069. Links to all molecular networking jobs run in GNPS have also been provided in the end of supporting information.

Results and discussion

A Thermo QE MS-based workflow for data acquisition parameter optimization

To reveal the variations of chemical compositions in different BRB extract products, two types of BRB extracts, liquid and powder samples, were obtained and evaluated in this study. Furthermore, data collected on a Thermo QE hybrid MS system using these samples were analyzed to highlight the importance of MS parameter optimization in establishing MNs for compound annotation. The schematic workflow of this study was shown in Figure 1, which can be divided into four modules: sample preparation, parameter setting comparison, data analysis and data visualization. First of all, two types of BRB samples were prepared for the detection of potentially different chemical profiles from these products. Next, 4 parameters in DDA acquisition were chosen for optimization. For discrete parameter, such as resolution, of which the options are limited by the instrument, all 4 resolutions provided were applied. For other parameters, since the values are continuous, we took the default/common value as a median, and a larger value and a smaller value were taken from both ends to observe their effects on the detection result. Collected raw data were uploaded to GNPS after conversion. The cosine score was generated by comparing it to the GNPS built-in database MS^2 graph. Following common practice [18], cosine score cut off was set at 0.7, which means that any feature with a matching score greater than 0.7 will be annotated, otherwise it will be used as an unknown feature to go to the next algorithm.

In the next step, annotated compounds serve as the core of MN, through the comparison with the MS^2 spectra of other unannotated features, to identify the adduct or isotope forms of the previously annotated compounds with the cosine score cut off at 0.7. In addition, GNPS built-in database also provide the reactions between compounds, through which reactants and products of chemical reactions can be linked. Based on the above algorithm, these features are successively linked together to form MNs with the core of annotated compounds. Figure S1 shows a cluster centered on kaempferol 7-O-glucoside in MN at and around m/z 449.101, in which the identified compounds were shown as blue dots, while the unidentified features were shown as gray dots. Centered with m/z 449.101, other compounds with a cosine score greater than 0.7 were annotated as its derivatives and connected. Then, each of these derivatives were centered, any features with a cosine score of more than 0.7 were connected to them. In practice, all identified compounds are seen as the centers of the network from primary nodes to secondary nodes. Either identified or unidentified features, connected after the first matching, are seen as the centers of the network from secondary nodes to other nodes. As the network spreads, additional compounds are connected into the MN to maximize the annotation potentials of detected spectra. Finally, quality indicators of these MNs are extracted to reflect the overall performance of the MNs. To simultaneously reflect multi-dimensional indicators, including nodes, pairs/edges, and networks, radar maps were used in our analysis for data visualization and parameter optimization. At the same time, some representative MNs are also displayed to verify the optimized results of the radar map.

Evaluating MN performance indicators under different experimental conditions

To generate simplified comparisons of each variable tested, radar maps with 8 analytical indicators were used to display the parameters and overall quality of the MNs that produced by different MS datasets. Among them, total nodes, identified nodes, nodes in network reflected the number of detected features in MS experiments; self-loop nodes, pairs/edges, cosine scores reflected the degree of correlation for these features; network and cluster coefficient reflected the complexity of MNs. In general, a higher value of the cosine score between the nodes indicates a higher similarity of the detected spectra, and a higher similarity of the chemical structure of two metabolites. Meanwhile, a higher clustering coefficient of a node in the molecular network indicates that it has a larger number of chemicals that share similar structures, and a higher influence to the molecular network can be represented by this node based on its similarity to other chemical structures. In this study, all cosine scores and clustering coefficients were reported as an average value of multiple LC-MS runs to represent the strong correlations of nodes in the molecular network and the robustness of the molecular network. In general, except self-loop nodes, larger values of other parameters indicate better MS spectra that are suitable for MN analysis. However, since the self-loop nodes and total nodes were positively correlated, the evaluation criterion was set as the larger the octagon produced by the radar map, the better the MS^2 spectrum quality under this parameter.

As a starting example, the radar maps of parameter optimizations using liquid BRB extract were shown in Figure 2. The detailed data which was used to draw the radar map is listed in Table S2. For resolution of MS^1, 17,500, 35,000, 70,000 and 140,000 were applied in this experiment. It is well-known that Q-Exactive mass spectrometer can achieve m/z 200 resolution up to 140,000 and less than 2 ppm accuracy to achieve reliable identification [19]. But too much resolution comes at the expense of reduced signal intensity [20]. As shown in Figure 2A, in positive mode, the MS^2 spectra of 70,000 resolution was significantly better as demonstrated by the largest octagon (yellow line). Compared to the second largest values, the identified nodes of MS^2 spectra of 70,000 resolution increased the most (with an 84.3% increase on average comparing to that of resolution at 35,000), and the cluster coefficients only increased 2.4% comparing to resolution at 17,500. While in negative mode (Figure 2E), the impact of resolution was not as dramatic as the positive mode. Compared to the second largest values at 35,000, the largest increase at 17,500 was only 18.9% in pairs/edges. Meanwhile, identified nodes and cosine score at 17,500 even decreased by 1.5% and 2.7% from the maximum at 35,000. In NCE optimization, gradient of 10, 40 and 70 were applied for the experiment. Different collision energies can result in the variations in fragmentation of ions, so the NCE value has a great potential to influence the MS^2 fragmentation patterns and therefore often call for careful optimization [21]. For NCE evaluation, NCE of 10 in both polarities was significantly better overall (Figure 2B and 2F). For both modes, almost all indicators except cosine scores and cluster coefficients reached to the maximum performance at NCE of 10. Most significantly, self-loop nodes increased about 230.9% in positive mode and networks increased about 196.9% in negative mode both compared to NCE of 40. The maximum of cluster coefficients was shown at NCE of 40 in both polarities, which was increased by 6.1% and 9.0% than those at NCE of 10 in positive and negative mode, respectively. As for the intensity threshold, 10E4, 10E5, and 10E6 were applied in the experiment. Intensity threshold refers to the MS^1 signal intensity threshold selected for the MS^2 spectrum generation. The intensity threshold is often used to remove the lower peaks of the signal because they may come from the noise and cannot generate meaningful MS^2 data [22]. As demonstrated in Figure 2C and 2G, the intensity threshold showed no significant difference between 10E4 and 10E5, while all indicators decreased significantly at 10E6 in both modes. In positive mode, the largest increase of 49.0 % appeared on cluster coefficients when comparing 10E5 to 10E4 (Figure 2C), while in negative mode, the largest increase of 46.5% shown up on self-loop nodes when comparing 10E4 to 10E5 (Figure 2G). Thus, 10E5 and 10E4 were recognized as the best options for positive and negative mode data collection, respectively. For exclusion time optimization, 10, 30 and 60 seconds were applied in the experiment. Exclusion time refers to the duration of dynamic exclusion in the selection of MS^1 signal for fragmentation. For example, if 10 seconds is selected, the same MS peak selected for MS^2 spectra generation in the top N list will not be fragmented for the second time within 10 seconds. The purpose of applying dynamic exclusion is to increase the compounds coverage by eliminating redundancy and allowing more precursor ions to be fragmented [23]. The effect of exclusion time on MNs was relatively weak compared with other parameters, with slightly better MS^2 performance in both polarities appearing at 60 s (Figure 2D and 2H). As the exclusion time increased, most indicators except cosine scores and cluster coefficients increased accordingly. For cosine scores, the maximum, appearing at 30 s in positive modes and 10 in negative mode, increased by 0.2% and 0.7% than 60 s. For cluster coefficients, the maximum in positive mode still showed up at 60 s in positive mode and 10 s in negative mode.

Similarly, the radar maps of parameter optimization for powder BRB extract were shown in Figure S2 with the detailed data listed in Table S3. Compared with BRB liquid sample, powder BRB extract showed some different trends when optimizing the tested parameters. With the increase of resolution in both polarities, the number of total nodes, nodes in network, self-loop nodes, pairs and networks kept decreasing (Figure S2A and S2E). Identified nodes showed a little different pattern by reaching the highest value at 35,000 (increased by 1.42% in positive mode and 20.5% in negative mode compared with that of 17,500). The average cosine score coefficient reached its maximum at 140,000 (increased by 2.7% compared with that of 17,500) in positive mode and 70,000 (increased by 1.4% compared with that of 17,500) in negative mode. The maximum value of the coefficient correspondingly appeared at 35,000 (increased by 20.3% compared with that of 17,500) and 17,500, respectively. Overall, the area of octagon decreased significantly as the resolution increase, and resolution of 17,500had the best MS^2 spectral performance in both positive and negative mode. Because NCE value strongly influencing the patterns of fragmentation ions, and these patterns are often different in the two polarities, the patterns in MN statistics between the two modes were deemed to be different (Figure S2B and S2F). Overall, according to the size of these octogens in radar maps, the best MN performance occurred at 70 at positive mode and 40 at negative mode, respectively. Similar to liquid BRB extract, in powder BRB extract, the performance of intensity thresholds was highly comparable between 10E4 and 10E5, which were both larger in octagon size comparing to 10E6 threshold (Figure S2C and S2G). The number of nodes, pairs and networks at 10E6 was significantly reduced and the precision and complexity of the MN kept similar (cluster coefficients in positive mode is only lower than that of 10E5 by 5.1%) or even increased (cosine score increased by 1.8% compared to 10E4 and cluster coefficient increased by 9.7% compared to 10E5 in negative mode). It can be inferred that most true MS^1 peak has a signal intensity greater than 10E5. When the intensity threshold was set at 10E6, some of the true MS^1 signals were not selected for fragmentation; meanwhile, those higher intensity peaks facilitated the formation of a MN with higher cosine scores and cluster coefficients. Although results at 10E6 showed more complex MN structure, that came at the expense of losing many of the nodes. Therefore, to achieve a balance between quality and quantity, 10E5 was the best setting for intensity threshold for both polarities in the powder BRB extract analysis. Exclusion time was still not a key factor in the powder BRB extract analysis, which is similar to the liquid BRB extract analysis in positive mode (Figure S2D and S2H). In terms of polarity, the negative mode performed better in MN performance than the positive mode under the same conditions for all parameters. The best parameters for both samples and polarities are summarized in Table S4.

In a previous study, the MNs generated by GNPS were also applied to optimize MS^2 mapping results, among which they only optimized the absolute threshold, and exclusion after n scans [6]. This research was performed with an Agilent mass spectrometer while our experiment was based on an Thermo QE instrument, which added reference values for parameter optimization for different mass spectrometer instruments. Their study showed similar trends of the impact from intensity threshold and exclusion after n scans but larger influence to their MN performance were observed comparing to ours, which could because that we were able to obtain an order of magnitude higher number of nodes than their results. In addition, after data collection, they conducted a secondary optimization by adjusting parameters of GNPS and Agilent data analysis software, which will be the focus of our follow-up studies. Another study focused on the parameter optimization, including MS^1 resolution, MS^2 resolution, data points per peak, for large protein [24]. Different from our study, this research was based on Q-TOFS and Orbitrap instruments and used DIA sampling methods. Eventually, the authors found that MS^1 resolution had little impact on DIA segments, and showed a trend of first rising and then falling, which are also different with our pattern of continuous reduction.

The impact of data acquisition parameters on the MN results

In order to visualize the influence of different parameter settings on the MNs generated in this study, the largest cluster in each MN was shown as examples in Figure 3 and Figure S3, for the liquid BRB extract and powder BRB extract, respectively. The color bars at the bottom of these figures were used to represent the intensities of nodes. Here, we found that in resolution, intensity threshold, and exclusion time groups, there was an interesting correlation between the peak intensity of MN nodes and the optimization results of radar maps: the parameters that perform well in the radar maps usually have deeper colors over many MN nodes, which represent higher intensities, in the overall cluster. But for the NCE, the correlation did not hold. We hypothesized that since the MN was produced by the combination of results under all parameters, intensity and radar map optimization results were expected to be consistent for compounds that were detectable in all selected parameter settings. Because the purpose of radar map was to screen out better quality MS^2 spectra, and higher quality spectra usually had higher intensity. However, for compounds that were not found in all parameter settings, especially the unique compounds that were only detected under specific parameters, radar maps could only reflect the quantity of MS^2 diagrams, but not their unique identities. Therefore, in Figure 4 and S4, the origins of reported MN nodes in these clusters (generally containing ~ 100 nodes) were shown in different colors. For example, in the first resolution column of Figure 4, compounds labeled black were detected at all four resolutions, while dark blue, red, yellow and green represented unique features detected at each tested value. Both liquid sample and powder sample in both polarities show very different origin distribution in NCE compared with other 3 parameters. For other 3 parameters, the three tested exclusion time for powder BRB extract in positive mode (Figure S4D) had as much as 94.4% in common nodes; and the three tested intensity thresholds for powder BRB extract in negative mode (Figure S4G) had the least common nodes (4%). The clusters in NCE group shown little common nodes (0~5%) but a lot of unique nodes (Figure 4B, 4F, S4B, S4F). In order to extrapolate our conclusions to the entire MN, all the nodes sources were counted in Figure 5 and Figure S5. Similar to the figures above, the color of the bar chart also represented the origin of nodes. Overall, resolution, intensity threshold, exclusion time had consistent trend of all feature numbers with total nodes in radar plots, that is, feature numbers decreased with the resolution or intensity threshold settings increased but increased with the exclusion time settings increased (Figure 5A, 5C, 5D, 5E, 5G, 5H, Figure S5A, S5C, S5D, S5E, S5G, S5H). Exclusion time had most common features (Figure 5D, 5H, Figure S5D, S5H), which changed from 61.6% (with exclusion time at 60 s for liquid BRB extract in positive mode) to 78.9% (with exclusion time at 10 s for liquid BRB extract in negative mode), and least unique features, which changed from 2.1% (with exclusion time at 30 s for liquid BRB extract in negative mode) to 7.1% (with exclusion time at 60 s for powder BRB extract in negative mode). Although resolution and intensity threshold did not have as many common features as exclusion time did, their common features were significantly larger than that of NCE (Figure 5B, 5F, Figure S5B, S5F). Nodes under the NCE parameter showed a very high degree of uniqueness, with NCE of 10 producing the largest number of unique compounds, accounting for about 80% of all the detected compounds with this setting. This finding, which was universal in both samples and both polarities, indicated that even the radar map showed very similar optimization results, different NCE values can generate totally different nodes for MN construction. The MS^2 results are pretty sensitive to NCE, but this may not be reflected by radar maps using MN indicators. Therefore, direct MNs inspections will help to overcome the bias of radar map evaluations when we need to focus not only on the number of MS^2, but also on the composition of those numbers, especially for those parameters shown fewer changed indicators while comparing different settings.

The identification result comparison of liquid and powder BRB extracts under both polarities

Even though MS^2 DDA can usually detect thousands of features / nodes in MNs, in untargeted metabolomics, the ones that get most attention are still those can be annotated. In GNPS, the MN is derived from the annotated compounds. In our study, the number of these annotated features were organized and compared in Figure S6. There was a total of 865 features identified. Overall, larger number of features were uniquely detected in positive mode than those in negative mode, in which the largest set of unique features (n=256) were detected in liquid sample under positive mode. Much larger number of overlap compounds in the same polarities (positive: n=78, negative: n=41) than that in the same sample (liquid: n=1, powder: n=3) were annotated, suggesting that the influence on the annotations from detection polarities were much larger than the two sample types we analyzed in this study. It is well known that BRB extracts are rich in polyphenols. According to previous literature of various berries studies[25], BRB Anthocyanin-Enriched Extracts ranges from 90 to 100 mg of gallic acid equivalents/g, which is much higher than the other kinds of berries, including blackberry, blueberry, cranberry, red raspberry, and strawberry. Therefore, polyphenol related features were picked from all identified features and organized in Table 1. Overall, 88 polyphenols and derivatives were detected. The table showed similar pattern as that of all features: more features in positive mode and more overlaps between the same polarities. Kaempferol and its derivatives were abundant and detected in both samples and polarities, while quercetin and its derivatives were more sensitive to polarity and most of them were detected in negative mode. There were less polyphenols difference by sample types, which is what we expected. A previous study detected polyphenols types and level in BRB wine produce via alcoholic fermentation. Seventeen poly phenols, 10 of which were also detected in our study, were found in the wine [26]. Some polyphenols, such as anthocyanins and tannins in BRB, are susceptible to degradation during food processing [27], which could explain the differences we observed between liquid and powder BRB extracts in our study. In chokeberries juice, small amounts of anthocyanins were present in large part in polymeric forms after 6-month storage at 25 °C [28]. But another study showed that total anthocyanins from berries can be well retained during long-term storage at −20 °C [29]. This contradictory information suggested that additional studies, such as our current study, are necessary to explore the polyphenol differences in BRB samples under different processing methods and storage conditions.

The MN comparison of liquid and powder BRB extracts under optimal parameters

After the optimal parameters of the two BRB extracts under both polarities were obtained from the radar maps, the raw data obtained using the optimal parameters in Table S4 were selected to generate the optimal MNs. Figure 6 showed the MNs formed by the best parameters of liquid and powder BRB extracts under positive mode analysis. The red dots indicated that these compounds detected from liquid BRB extract, the blue dots indicated that these compounds detected from powder BRB extract, and dots with mixed colors suggested these compounds can be detected from both type of samples, with the ratio of red to blue represented the intensity ratio of the compound under the optimal parameters. The optimal parameters of the two products are not identical, and the MNs generated under the optimal parameters were also quite different. More self-loop nodes came from liquid BRB, while nodes from powder tends to cluster into networks, which highlighted the possible influences of food processing steps to phytochemicals within these dietary supplements. In addition, many of the compounds in the figure were derived entirely from one sample, indicating that a good proportion of compounds in different BRB extracts vary under different processing conditions. Two representative polyphenols in positive mode and another two polyphenols in negative mode were recognized and highlighted/labeled in Figure 6. All of them can only be detected from BRB powder. In positive mode, kaempferol 3-glucuronide and cyanidin 3-galactoside were in the largest cluster, while kaempferol-3-glucoside-6-p-coumaroyl and quercetin-3-o-deoxyhexoside were self-loop in negative mode, which implicated that many polyphenol derivatives may be contained and connected with each other in BRB powder extract in positive mode. It is acknowledged that even for the selected samples, our specific type of instrument and the selected testing conditions reported in this study, we can be confident that this set of parameters was optimized, it is important to note that our optimal parameters here have limited applicability for other types of samples, instruments, or parameters. However, by showing different optimal parameters for liquid and powder RBR extracts, we demonstrated that our optimization process is capable of selecting a relatively good experimental parameter set, and that our optimization process is necessary to obtain higher quality MS^2 spectra.

Conclusion

In this study, we demonstrated that parameter optimization for DDA in untargeted metabolomics analysis is necessary to obtain MS^2 spectra of higher quality and quantity, which are essential for building large scale MNs for better compound annotation. GNPS platform provides a new and effective way for the identification of unknown compounds, and in this study, we aim to identify and optimize a set of experimental/data analysis parameters and their corresponding performance indicators of molecular network based on GNPS, which were able to also provide a quantitative basis for the determination of the quantity and quality of MS2 spectrum. In turn, sufficient and higher quality MS2 spectrum is a prerequisite for further improving the capability of GNPS platform identification. Therefore, the optimization of experimental parameters is essential in the identification of unknown metabolites based on GNPS platform. Among 4 parameters tested in this study, resolution and intensity threshold were optimized for enhanced MS^2 signal intensity, while the optimization of NCE could increase the spectra diversity and enable the annotations of many unique compounds based on their fragmentation patterns. Resolution, NCE and intensity threshold have greater influences on MS^2 spectra quality, while exclusion time have less influence. By optimizing the parameters of the two types of BRB extracts separately, we noted that the parameter optimization can be sample type-specific and MS polarity dependent. In summary, we believe our study provided a Thermo QE mass spectrometer-based workflow for improving the integration of MS^2 spectra quality and MN construction for better small molecule detection and annotation, and also provided references for optimizing parameters in similar analytical conditions. Moving forward, as we acknowledge that this study only had a small coverage of sample types and several major MS parameters optimized, we plan to continue this line of work to other natural products and small molecule analysis with extended optimizations of other MS parameter as well as post-data collection analysis to fill more gaps of knowledge in the complicated and challenging field of untargeted metabolomics analysis.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Grant Support:

Research reported in this publication was supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number R35GM133510.

Reference

1. Dettmer K, Aronov PA and Hammock BD, Mass spectrometry-based metabolomics, Mass spectrometry reviews, 2007, 26, 51–78 [PubMed: 16921475]
2. Schrimpe-Rutledge AC, et al. , Untargeted metabolomics strategies—challenges and emerging directions, Journal of the American Society for Mass Spectrometry, 2016, 27, 1897–1905 [PubMed: 27624161]
3. Guo J and Huan T, Comparison of Full-Scan, Data-Dependent, and Data-Independent Acquisition Modes in Liquid Chromatography–Mass Spectrometry Based Untargeted Metabolomics, Analytical Chemistry, 2020, 92, 8072–8080 [PubMed: 32401506]
4. Zhao X, et al. , Comprehensive strategy to construct in-house database for accurate and batch identification of small molecular metabolites, Analytical chemistry, 2018, 90, 7635–7643 [PubMed: 29807420]
5. Barbier Saint Hilaire P, et al. , Comparative Evaluation of Data Dependent and Data Independent Acquisition Workflows Implemented on an Orbitrap Fusion for Untargeted Metabolomics, Metabolites, 2020, 10, 158
6. Olivon F, et al. , Optimized experimental workflow for tandem mass spectrometry molecular networking in metabolomics, Analytical and bioanalytical chemistry, 2017, 409, 5767–5778 [PubMed: 28762069]

7. Jeffryes JG, et al. , MINEs: open access databases of computationally predicted enzyme promiscuity products for untargeted metabolomics, Journal of cheminformatics, 2015, 7, 44 [PubMed: 26322134]

8. Wang M, et al. , Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking, Nature biotechnology, 2016, 34, 828–837

9. Yang JY, et al. , Molecular networking as a dereplication strategy, Journal of natural products, 2013, 76, 1686–1699 [PubMed: 24025162]

10. Soares V, et al. , Extending compound identification for molecular network using the LipidXplorer database independent method: A proof of concept using glycoalkaloids from Solanum pseudoquina A. St.-Hil, Phytochemical Analysis, 2019, 30, 132–138 [PubMed: 30328225]

11. Santos AL, et al. , Identification of flavonoid-3-O-glycosides from leaves of Casearia arborea (Salicaceae) by UHPLC-DAD-ESI-HRMS/MS combined with molecular networking and NMR, Phytochemical Analysis, 2021

12. Newman ME, The structure and function of complex networks, SIAM review, 2003, 45, 167–256

13. Elie N, Santerre C and Touboul D, Generation of a Molecular Network from Electron Ionization Mass Spectrometry Data by Combining MZmine2 and MetGem Software, Analytical chemistry, 2019, 91, 11489–11492 [PubMed: 31429549]

14. Aron AT, et al. , Reproducible molecular networking of untargeted mass spectrometry data using GNPS, Nature protocols, 2020, 15, 1954–1991 [PubMed: 32405051]

15. Smoot ME, et al. , Cytoscape 2.8: new features for data integration and network visualization, Bioinformatics, 2011, 27, 431–432 [PubMed: 21149340]

16. Aguilar-Mogas A, et al. , imet: A network-based computational tool to assist in the annotation of metabolites from tandem mass spectra, Analytical Chemistry, 2017, 89, 3474–3482 [PubMed: 28221024]

17. Wagner A and Fell DA, The small world inside large metabolic networks, Proceedings of the Royal Society of London. Series B: Biological Sciences, 2001, 268, 1803–1810 [PubMed: 11522199]

18. Yang R, et al. , Chemical composition and pharmacological mechanism of Qingfei Paidu Decoction and Ma Xing Shi Gan Decoction against Coronavirus Disease 2019 (COVID-19): in silico and experimental study, Pharmacological Research, 2020, 104820 [PubMed: 32360484]

19. Michalski A, et al. , Mass spectrometry-based proteomics using Q Exactive, a high-performance benchtop quadrupole Orbitrap mass spectrometer, Molecular & Cellular Proteomics, 2011, 10

20. Yu T, et al. , apLCMS—adaptive processing of high-resolution LC/MS data, Bioinformatics, 2009, 25, 1930–1936 [PubMed: 19414529]

21. Hartel NG, Liu CZ and Graham NA, Improved discrimination of asymmetric and symmetric arginine dimethylation by optimization of the normalized collision energy in LC-MS proteomics, Journal of Proteome Research, 2020

22. Schuhmann K, et al. , Intensity-independent noise filtering in FT MS and FT MS/MS spectra for shotgun lipidomics, Analytical chemistry, 2017, 89, 7046–7052 [PubMed: 28570056]

23. Zhang Y, et al. , Effect of dynamic exclusion duration on spectral count based quantitative proteomics, Analytical chemistry, 2009, 81, 6317–6326 [PubMed: 19586016]

24. Bruderer R, et al. , Optimization of experimental parameters in data-independent mass spectrometry significantly increases depth and reproducibility of results, Molecular & Cellular Proteomics, 2017, 16, 2296–2309 [PubMed: 29070702]

25. Ma H, et al. , Evaluation of polyphenol anthocyanin-enriched extracts of blackberry, black raspberry, blueberry, cranberry, red raspberry, and strawberry for free radical scavenging, reactive carbonyl species trapping, anti-glycation, anti-β-amyloid aggregation, and microglial neuroprotective effects, International journal of molecular sciences, 2018, 19, 461

26. Lim JW, Hwang HJ and Shin CS, Polyphenol compounds and anti-inflammatory activities of Korean black raspberry (Rubus coreanus Miquel) wines produced from juice supplemented with pulp and seed, Journal of agricultural and food chemistry, 2012, 60, 5121–5127 [PubMed: 22563950]

27. Howard LR, et al. , Processing and storage effect on berry polyphenols: challenges and implications for bioactive properties, Journal of agricultural and food chemistry, 2012, 60, 6678–6693 [PubMed: 22243517]

28. Wilkes K, et al. , Changes in chokeberry (Aronia melanocarpa L.) polyphenols during juice processing and storage, Journal of agricultural and food chemistry, 2014, 62, 4018–4025 [PubMed: 24274724]

29. Hager A, et al. , Processing and storage effects on monomeric anthocyanins, percent polymeric color, and antioxidant capacity of processed black raspberry products, Journal of Food Science, 2008, 73, H134–H140 [PubMed: 19241590]

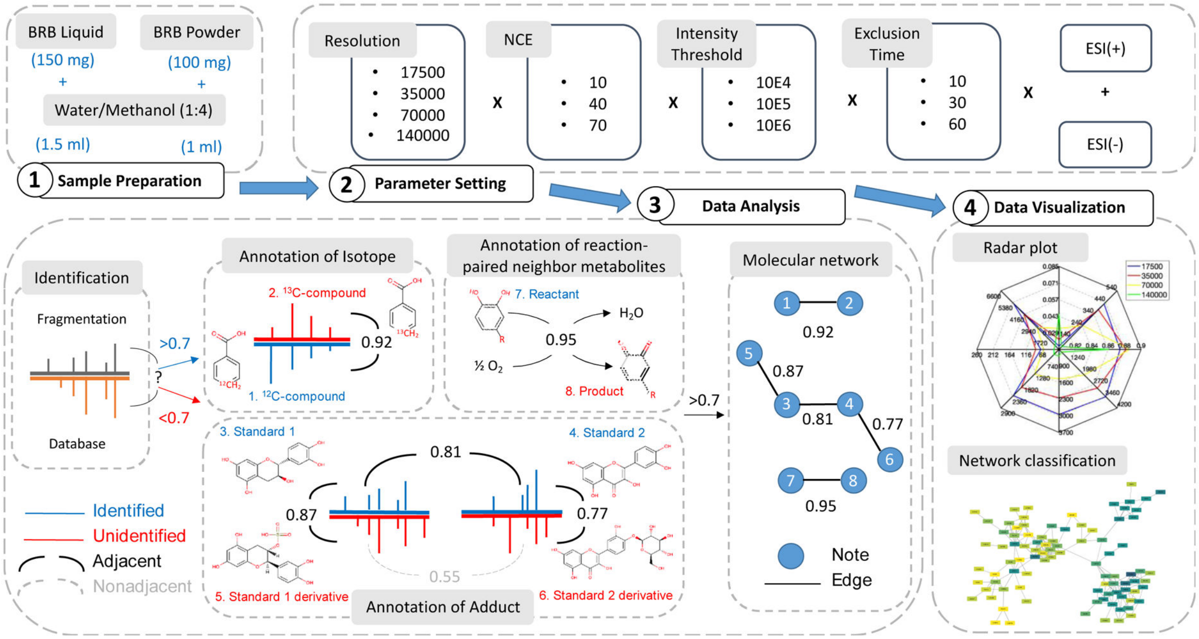

Figure 1.

The workflow of our molecular networking-assisted metabolomics study. 1. BRB liquid and powder samples were prepared. 2. Resolution, NCE, intensity threshold and exclusion time were divided into gradients and cross-combined for LC-MS detection. 3. Detected data were used to build molecular network. 4. Indicators for molecular network were summarized to assess the performance of different parameter settings.

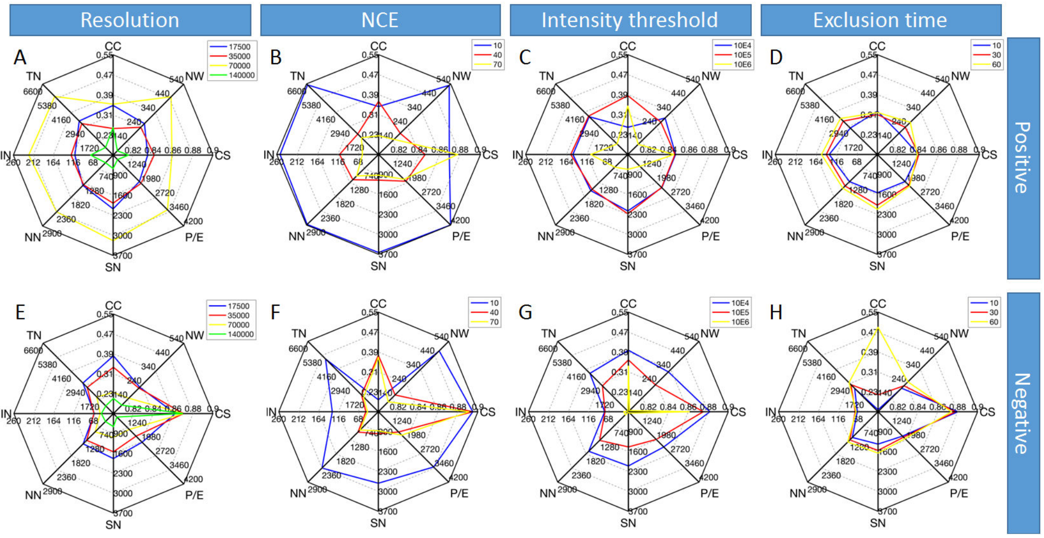

Figure 2.

Average statistical values for molecular network with liquid BRB extract raw data generated under different resolution, NCE, intensity threshold, and exclusion time. The radar plot matrix is divided into two rows and four columns, and the intersection point of each row and column is the radar plot of a certain parameter under the certain experimental condition. The radar plot has eight indicators in eight directions, and the line segments connect the eight values to form an octagon. TN: total nodes; IN: identified nodes; NN: nodes in network; SN: self-loop nodes; P/E: pairs/edges; CS: cosine scores; NW: networks; CC: cluster coefficients. Colors represent the different gradients of the parameters.

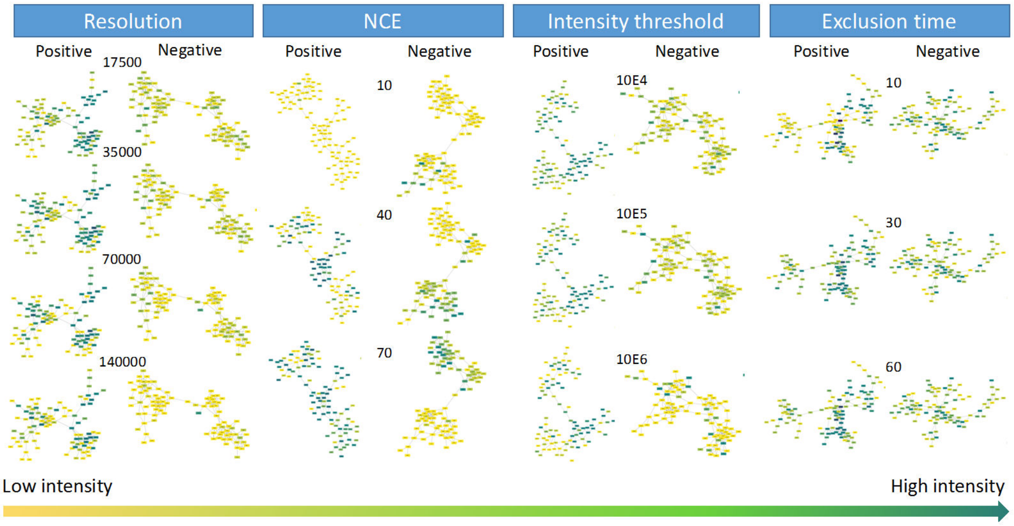

Figure 3.

The intensity of the largest cluster in each MN under different parameter settings using liquid BRB extract. Colors represent the signal intensity of the node. The darker the color, the stronger the signal intensity. The color distribution in these clusters represents the relationship between signal intensity of nodes and parameter changes.

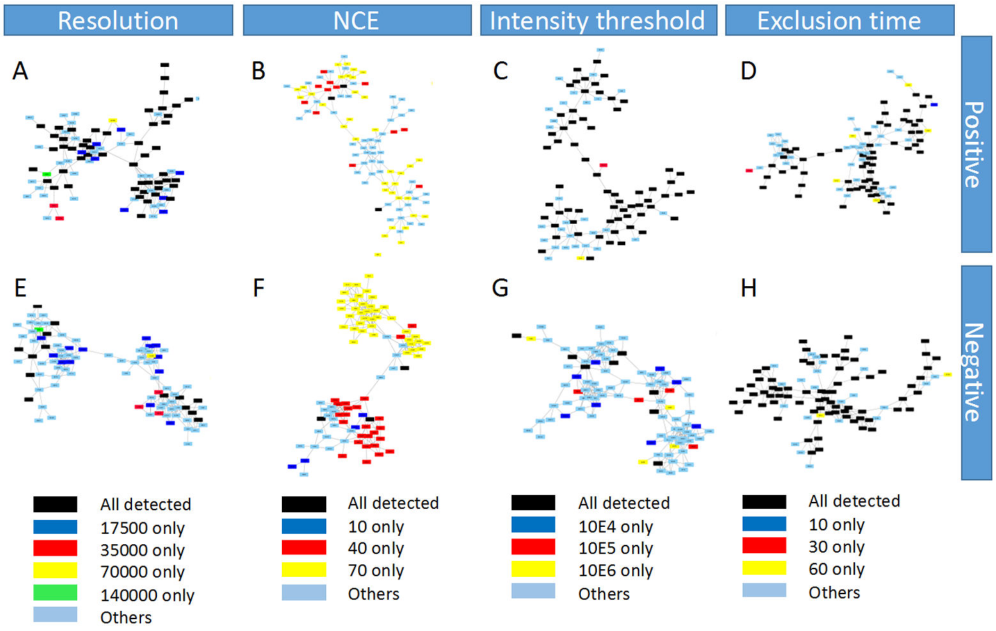

Figure 4.

The node origins of the largest cluster of each MN under different parameter settings using liquid BRB extract. Colors represent the source of nodes.

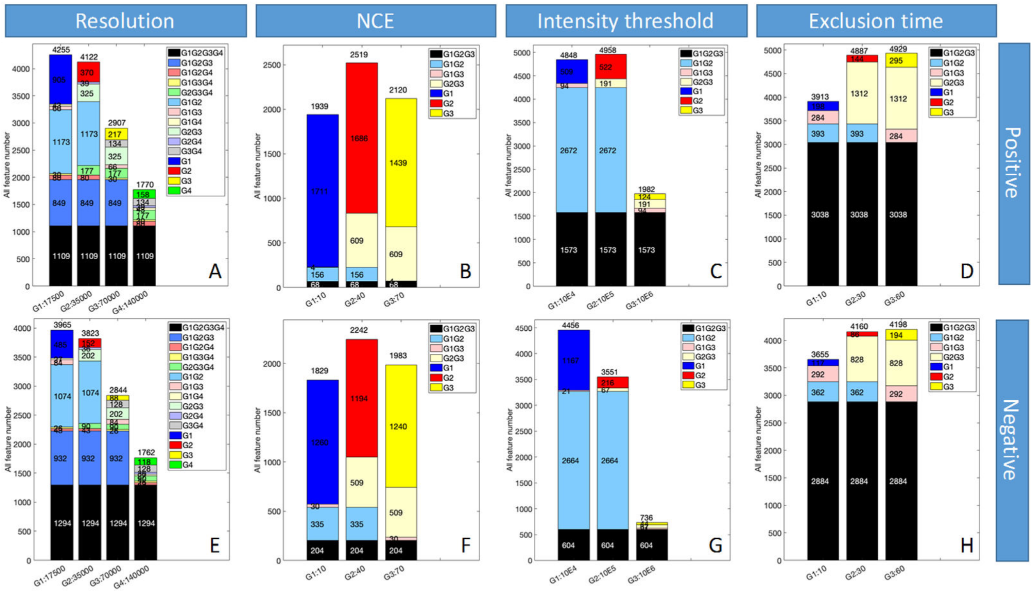

Figure 5.

The node origins distribution of each MN under different parameter settings using liquid BRB extract. Colors represent the source of nodes.

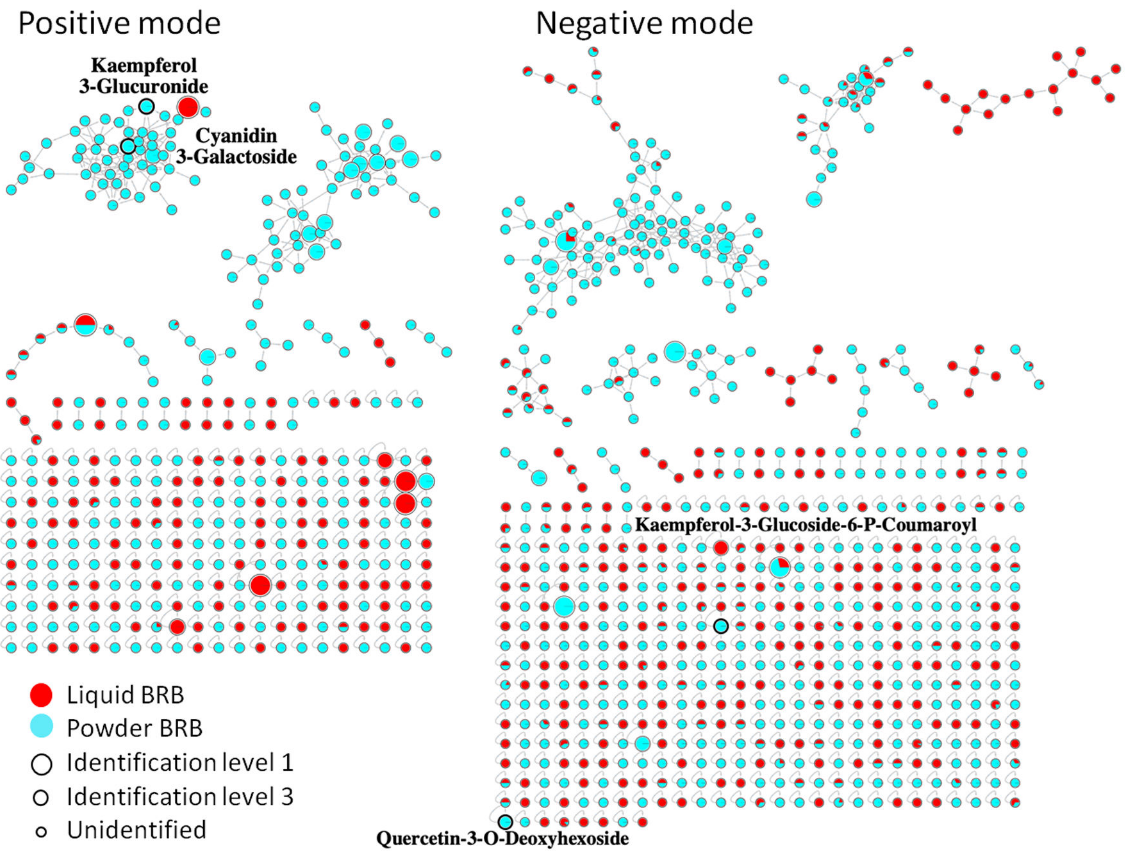

Figure 6.

The overview molecular network under optimized detection condition for the comparison of liquid and powder BRB extract samples. Colors represent the source of nodes. Size represents the identification level in GNPS. Bold black circles highlight identified polyphenol derivatives with their annotation aside.

Table 1.

The list of detected polyphenols in this study.

Number	Compound	Molecular weight	Liquid positive	Liquid negative	Powder positive	Powder negative
1	Kaempferol 3-Glucuronide	462.079	+	+	+	+
2	Kaempferol-3-Glucoside-6-P-Coumaroyl	594.137	+	+	+	+
3	Kaempferol-3-O-Glucoside	448.100	+	+	+	+
4	Kaempferol-3-O-Rutinoside	594.158	+	+	+	+
5	Quercetin-3-O-Beta-Glucopyranoside	465.388	+	+	+	+
6	Rutin	610.153	+	+	+	+
7	Luteolin 4'-O-Glucoside	448.100	+	+	+	
8	Quercetin-3-O-Alpha-L-Rhamnopyranoside	448.100	+	+	+	
9	Kaempferol 7-Neohesperidoside	594.158	+	+		
10	Kaempferol 7-O-Glucoside	448.100	+		+	+
11	Quercetin-3-Glucuronide	478.075	+		+	+
12	Rhoifolin	578.163	+		+	+
13	Syringaresinol	440.144	+		+	+
14	Aloenin	432.103	+		+	
15	Beta-Sitosterol	396.592	+		+	
16	Kaempferol-3-Glucoside-3-Rhamnoside	594.158	+		+	
17	Neohesperidin	610.522	+		+	
18	Quercetin-3-O-Alpha-L-Rhamnopyranoside	448.100	+		+	
19	Quercetin-3,4-O-Di-Beta-Glucopyranoside	625.992	+		+	
20	Stigmastanol	398.492	+		+	
21	Guajavarin	434.084	+			+
22	Abruquinone B	428.087	+			
23	Benzoic Acid Derivative 1	469.184	+			
24	Benzoic Acid Derivative 2	412.164	+			
25	Carpachromene	358.081	+			
26	Cyanidin 3-Galactoside	448.104	+			
27	Hesperidin	610.190	+			
28	Isoquercetin-3-O-Alpha-L-Rhamnopyranoside	448.100	+			

Number	Compound	Molecular weight	Liquid positive	Liquid negative	Powder positive	Powder negative
29	Isorhamnetin-3-Rutinoside	624.169	+			
30	Kaempferol-3-O-Alpha-L-Rhamnopyranosyl(1-2)-Beta-D-Glucopyranoside-7-O-Alpha-L-Rhamnopyranoside	740.662	+			
31	Kaempferol-3-O-B-Glucoside-7-O-A-Rhamnoside	594.172	+			
32	Matairesinol	375.167	+			
33	Myricetin-3-Galactoside	480.090	+			
34	Myricetin-O-Hexosyl-Deoxyhexoside	626.147	+			
35	Peonidin 3-O-Glucoside	462.115	+			
36	Quercetin-3-O-Beta-D-Galactoside	464.095	+			
37	Quercetin-3-Rhamnoside-7-Glucoside	610.148	+			
38	Rosmarinic Acid	398.040	+			
39	Hyperoside	464.095		+	+	
40	Kaempferol-7-O-Hexosyl(1-2)Deoxyhexoside	594.158		+	+	
41	5,6,2'-Trimethoxyflavone	312.100		+		+
42	Ellagic Acid	302.007		+		+
43	Naringenin-7-O-Glucoside	434.092		+		+
44	Quercetin	302.043		+		+
45	Quercetin 3-O-Glucuronide	478.074		+		+
46	Quercetin-3-O-Deoxyhexoside	446.079		+		+
47	Quercetin-3-O-Deoxyhexosyl(1-2)Deoxyhexoside	594.154		+		+
48	Quercetin-3-O-Hexosyl-Deoxyhexoside	610.152		+		+
49	Zapotin	342.111		+		+
50	Adrenaline Bitartrate	333.107		+		
51	Apigenin-7-O-Glucuronide	446.086		+		
52	Epigallocatechin Gallate	459.378		+		
53	Iridin	522.137		+		
54	Isotectorigenin, 7-Methyl Ether	328.095		+		
55	Kaempferol-3-O-Hexosyl-Deoxyhexoside	594.154		+		
56	Maesopsin	288.064		+		
57	Myricetin-3-O-Hexosyl(1-2)Deoxyhexoside	626.121		+		
58	Quercetin 3-O-Neohesperidoside	610.162		+		
59	Quercetin-4-Glucoside	462.080		+		

Number	Compound	Molecular weight	Liquid positive	Liquid negative	Powder positive	Powder negative
60	Diosmin	608.173			+	+
61	Epicatechin Gallate	442.089			+	+
62	Isoquercitin	464.095			+	+
63	Procyanidin B1	578.148			+	+
64	Avicularin	434.084			+	
65	Calycosin-7-O-Beta-D-Glucoside	430.126			+	
66	Cinnamtannin A3	1442.332			+	
67	Cyanidin-3-O-Galactoside	448.100			+	
68	Isorhamnetin-3-O-Rutinoside	624.092			+	
69	Kaempferol-3-Rhamnoside-4-Rhamnoside-7-Rhamnoside	724.221			+	
70	Kaempferol-7-Neohesperidoside	594.158			+	
71	Methoxy-Quercetin-O-Hexosyl-Dideoxyhexoside	770.229			+	
72	Procyanidin A1	576.126			+	
73	Procyanidin A2	576.126			+	
74	Procyanidin B2	578.099			+	
75	Quercetin 3-O-Malonylglucoside	550.095			+	
76	Quercetin-3-O-Robinobioside	610.151			+	
77	3,4-Di-O-Caffeoylquinic Acid	554.082				+
78	3,4-Dicaffeoylquinic Acid	514.112				+
79	Catechin Gallate	442.098				+
80	Chlorogenic Acid	354.096				+
81	Cyanidin-3-O-Alpha-Arabinopyranoside	416.074				+
82	Delphinidin-3-Rutinoside	609.992				+
83	Epicatechin	290.078				+
84	Eriodictyol-7-O-Glucoside	450.092				+
85	Kaempferol-7-O-Hexoside	448.097				+
86	Luteolin-7-Glucoside	448.101				+
87	Oleuropein	540.185				+
88	Quercetin 3-(6'-Acetylglucoside)	506.108				+