Imperial College London

Imperial College London

Faculty of Medicine

Department of Metabolism, Digestion and Reproduction

Molecular networking for the fusion of untargeted LC-MS metabolomic datasets

Author: Chang Liu

Supervised by: Dr Rui Pinto, Dr Tim Ebbels

Submitted in partial fulfilment of the requirements for the MRes degree in Biomedical Research (Data science) of Imperial College London

August 2021

Statement of Originality

I certify that this thesis, and the research to which it refers, are the product of my own work, conducted during the current year of the MRes in Biomedical Research at Imperial College London. Any ideas or quotations from the work of other people, published or otherwise, or from my own previous work are fully acknowledged in accordance with the standard referencing practices of the discipline.

The datasets and annotations are provided by Dr Rui Pinto.

Abstract

Background: Integrating results of different experiments can enhance the power of metabolomics studies. When acquiring untargeted data by Liquid chromatography - mass spectrometry (LC-MS), the majority of the metabolomic features are not annotated, thus it is hard to match them across datasets using their chemical identities. Usually, retention time (RT), mass-to-charge ratio (m/z) and feature intensity (FI) are the only information characterising each feature, and can be used for feature matching. Though similar, the same feature has different RT, m/z and FI in different datasets, which makes it a challenge to compare and combine them.

Methods: To address this issue, we introduce a strategy that uses molecular networking and using only RT, m/z and FI values of features. Intra-dataset networks are constructed by matching isotopologues and adducts, and inter-dataset shifts in RT, m/z and FI are calculated accordingly. Additionally, we explored a strategy to evaluate the matching performance without using annotations.

Results: In our main example, 2465 feature matches were found from 6327 and 10910 features. There are 567 annotations available, in which 527 (92.9%) were found correctly.

The code is available at https://github.com/ChangLiuuczlcl8/Untargeted-LC-MS-Metabolomic.

1. Introduction

1.1. Metabolomics

Metabolomics is the study of products of cell metabolism (metabolites) within cells, biofluids, tissues or organisms, which are influenced by both genetic and environmental factors. Metabolomics is a non-invasive approach that is closely linked to the phenotype and has been used widely in biomedical research, e.g., biomarker discovery (Xia et al., 2013) and drug safety screens (Wishart, 2008a).

Metabolomics experiments can be classified as targeted or untargeted. Targeted metabolomics focuses on the analysis of selected molecules and requires prior knowledge of metabolites of interest. It is used when a specific biochemical question needs to be answered. Untargeted metabolomics is usually hypothesis-generating, thus instead of selecting specific metabolites beforehand, it aims at measuring the maximum number of metabolites as possible from biological samples without any intended bias (Schrimpe-Rutledge et al., 2016). Both targeted and untargeted metabolomics studies follow a similar pipeline: experimental design, sample collection, sample preparation, sample analysis, data processing, and data analysis (Dettmer et al., 2007).

Mass spectrometry (MS) and nuclear magnetic resonance (NMR) are the two most common analytical techniques to generate untargeted metabolomics data. NMR analyses molecular structure by observing and measuring the interaction of nuclear spins, and MS measures metabolites according to the mass-to-charge ratio of their ions. Comparing with NMR which has very high reproducibility but low sensitivity, MS is much more sensitive but with only average reproducibility (Wishart, 2008b, Zhou et al., 2012).

For analysis using MS, metabolites can be directly injected into the mass spectrometer or through a coupled chromatographic system. LC-MS is a commonly used high throughput combination to further separate metabolite classes, as these elute at different times according to their chemical and physical properties. It reduces the complexity of individual spectra by providing an additional dimension of retention time, therefore it is useful when analysing complex mixtures. The typical structure of an untargeted metabolomics dataset is a feature matrix where each row is a sample, and each column is a feature. A feature is typically a peak or signal that represents a chemical compound. In our context, the noise refers to the artefact variables that are wrongly created from high baseline areas, impurities, and instruments (e.g., polymers from chromatographic column, sample containers, solvents, etc).

Comparing and combining results between experiments is important especially in large-scale epidemiological and clinical research studies (Lewis et al., 2016), as increasing the size of datasets increases the power of a study. However, the detected metabolomic features are not easy to be annotated automatically by a tentative metabolite identification. Different datasets have different retention times, similar but not equal mass to charge ratios, and retention time is not a predictable property of the molecule, as it depends on the sample and cohort. Additionally, many metabolites have the same mass and come at similar retention times. Besides, metabolite databases are always considered incomplete. Comparing to proteomics, known protein sequences and enzyme cleavage patterns can predict peptide sequences and fragmentation spectra, fragments are relatively unpredictable in metabolomic studies (Schrimpe-Rutledge et al., 2016). Annotating untargeted metabolomics datasets is a task for specialised analysts and very time-consuming, thus the annotations of most metabolites are not available for all features. RT, m/z and FI are the only information that can be used to match two features with the same identification across datasets. This makes it a challenge to compare and combine different datasets.

1.2. Previous work

Related works can mainly be divided into two types: matching features across samples, and matching features across datasets. Finding correspondence of features between samples is not exactly the same as between datasets. In the first case, the raw data is used, containing both rich information at the m/z level as well as retention time profiles, while in the second case, the data is limited as it has already been peak-picked and exists in a tabular format where each feature is represented only by RT, m/z and FI.

Matching features across samples has allowed metabolomics its current format. Many software packages exist (XCMS (Smith et al., 2006), MZmine2 (Pluskal et al., 2010), MS-DIAL (Tsugawa et al., 2015)) which find peak correspondence across samples using m/z and retention time profiles from the raw data.

A good review of the feature matching across datasets has been made by Åberg et al. (2009), where they analyse the correspondence problem, discuss current state-of-the-art methods for synchronising samples, and predict the properties of future methods. Our work has similarities to CAMERA (Kuhl et al., 2012). CAMERA is used to extract compound spectra, annotate isotopologue and adduct peaks, and propose an accurate compound mass. They also use raw data (not only RT, m/z, FI values) including retention time profiles to find the isotopologues and adducts of the same metabolite. Our work also relates to the approaches by network strategy (GNPS (Wang et al., 2016)), though they apply it to tandem mass spectrometry (MS/MS) spectra thus taking advantage of fragmentation, while we apply it to median values of m/z across a dataset. GNPS groups sets of spectra from molecular families even when the spectra themselves are not identified and represents them as molecular networks with spectrums as nodes and spectrum-to-spectrum alignments as edges. The toolbox M2S (Pinto et al., 2021) defines feature correspondence between two similar experiments by finding their shifts in m/z, RT and FI by doing one-to-one feature match. This toolbox only requires the values of RT, m/z and FI, and does not require previously annotated feature anchors. It does one-to-one feature matching and thus it may struggle to model the inter-dataset shift when the retention times are very dissimilar. A similar method to ours (Habra et al., 2021), first bins similar m/z features at close retention times, then finds "anchor" matches one-by-one using the features of higher intensity only. This method aligns only the RT dimension and can only work if the FI of matches are correlated.

1.3. Aim and hypothesis

The method described by Pinto et al. (2021) in toolbox M2S will be used as the baseline of our method. The overall objective of this project is to use the isotopologues and adducts to find the shifts in m/z, RT and FI between two datasets to better match metabolomics features across datasets than the baseline method. Comparing to only using one-to-one matches, our method matches the features in a more robust way by finding the intra-dataset matching of isotopologues and adducts prior to inter-dataset feature matching, which increases the quality of the modelling of the inter-dataset shifts by reducing the number of false-positive matches. This allows datasets with larger differences to be combined. The overall aim will be divided into 3 sub-aims:

1. Using the isotopologues and adducts to construct intra-dataset metabolite subnetworks. This includes matching of isotopologues, defining appropriate thresholds, and matching of adducts.

2. Matching metabolite subnetworks across datasets and using the results to improve the feature matching process of the baseline model. This includes matching of metabolite subnetworks, deleting multiple matches, and deleting poor matches.

3. Evaluating the performance of the new matching algorithm.

2. Materials and Methods

2.1. Sample preparation and data acquisition

This work was performed by the National Phenome Centre at Imperial College London and by Metabometrix. Serum (MESA datasets) and plasma lithium heparin samples (Airwave dataset) were used. UPLC-MS profiling analysis for lipids with electrospray ionisation in the positive mode (ESI+) was performed as previously described (Izzi-Engbeaya et al., 2018, Lewis et al., 2016, Dona et al., 2014). The same procedure was used in the preparation of the Airwave plasma and MESA serum samples, except that 100µl of samples was used without dilution prior to addition of isopropanol for the lipidomics analyses. All analyses were acquired on Acquity UPLC systems coupled to Xevo G2-S. There were small differences in the composition of the mobile phases, thus all datasets are expected to have similar adduct composition. The mobile phase gradients and flow rate were similar in Airwave and MESA phase 2, thus the retention times of features are expected to be comparable along their 12 minutes. Those gradients and flow rates were very different in Mesa phase 1, in which a chromatogram was acquired in only 6 minutes.

Data processing

This work was performed by the National Phenome Centre at Imperial College London and Dr Rui Pinto. Peak picking was completed using Bioconductor R-package XCMS (Smith et al., 2006). Briefly, peaks were picked using “centWave” method using the following parameters for UPLC MS ESI+: 15 ppm tolerance, peak width (8, 20), signal-to-noise threshold (“snthresh”) 10, noise level 300 and prefilter (6, 1000); After peak grouping, non-linear retention time correction was applied, then missing values were imputed which yielded a table of samples (rows) by features (columns).

Drift correction was done using a method previously described (Dunn et al., 2011). Negative values were replaced with zeros, and the data were natural log-transformed after adding one. We filtered the data based on retention time to exclude non-retained features – only features between retention times 0.45-12 (in minutes) were accepted. We used principal components analysis (PCA) to identify samples that were outliers and excluded them. Further, we excluded values that were >5 median absolute deviations (MAD) from the median. Finally, we transformed the data into pseudo z-scores using median and MAD.

2.1.3. UPLC-MS metabolite annotation

This work has been performed by the National Phenome Centre and Dr Gonçalo Graça. Lipid annotation was initially completed by matching accurate mass fragmentation measurements to reference spectra from online databases (LIPID MAPS (Fahy et al., 2007), Metlin, HMDB) and previous publications. Where chemical reference materials were commercially available (Avanti Polar Lipids, Sigma Aldrich, Cayman Scientific), they were used to generate definitive molecular identification by direct matching of chromatographic and spectral qualities (including accurate mass, MS/MS spectra, and isotopic distribution) to those observed in the profiling data.

2.3. Workflow

The workflow is presented in Figure 1. The input of this method is two tables which contain feature sets (m/z, RT, FI) as their columns. The output is a table, each row of which contains the information for each feature match (see Supplementary Table 3.). There are optional inputs for parameters and will be described in detail in each step. Two feature sets are referred to as reference and target in this paper. Feature difference is calculated by target feature minus reference feature, and figures are made in relation to the reference dataset.

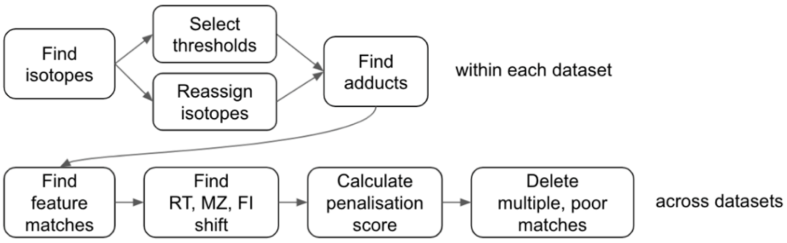


Figure 1. Method workflow. Firstly, isotopologue matching and adduct matching are performed within each dataset. isotopologues are detected and reassigned. Thresholds selection is an optional process. Adducts are matched within each layer of isotopologues. Then feature matching is performed across datasets. We find all feature matches and use them to calculate the trend shift of RT, m/z and FI. Penalisation scores are calculated using standardised residuals from the shift. Multiple and poor matches are deleted according to the penalisation score.

The essential assumption of this method is that the retention times of the same metabolomic features in both datasets are positively correlated, although there could be a non-linear shift between them. Additionally, within a dataset, all features corresponding to a molecule (isotopologues, adducts, fragments) are expected to elute at the same retention time and be highly correlated, and in some cases, their m/z difference is known.

2.3.1. Intra-dataset isotopologue matching

The objective in this section is to find pairs of features that may be isotopologues of each other, as well as to define thresholds for RT, m/z and correlation when matching different entities. Two atoms are called isotopes if they have the same number of protons and different number of neutrons in their nuclei. Isotopologues are molecules that only differ in their isotopic composition (McNaught and Wilkinson, 1997), which means at least one atom has a different number of neutrons. While isotopologues have similar chemical properties including identical RTs in LC-MS data, they have different masses, hence different m/z values in mass spectrometry.

In practice, a pair of features is defined to be isotopologues if they have the same RT, high correlation, and m/z difference equal to the mass of a neutron. Here, this is an experimental value obtained from previous analyses, defined as the mode of the distribution of mass differences between putative isotopologues, with the value 1.003355 Da (the mass of a neutron is 1.00866491588 Da). Additionally, the correlations are calculated across samples in four ways using different percentiles: 1 using all samples; 2 using only lowest 50% intensity samples; 3 using only 50% highest; 4 using 25 to 75th percentiles; and the highest one is chosen among them. Thus, pairs of features abiding by those definitions within specific thresholds (absolute RT difference < 0.005 minutes, Pearson correlation across samples > 0.7, absolute m/z difference < 0.005 Da) are considered to be isotopologues in the default setting. These three thresholds and m/z difference for isotopologues are changeable parameters adapted to each dataset.

After the initial isotopologue matches are found, a disconnected network is built within the dataset. Each connected component is a group of isotopologues, and the isotopic class (M+0, M+1, etc) for each feature is decided by the relative m/z value in the group (see Figure 2). M+0 is the monoisotopic mass, and M+1 is the isotopologue that has one atom with one more neutron. For the cases where there exist wrongly matched isotopologue pairs see Figure 14 in the Supplementary method.

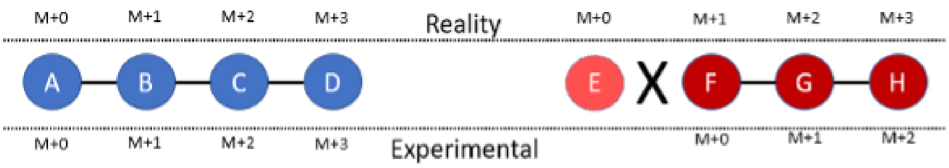


Figure 2. Two connected components of isotopologues. In blue, isotopologues A is the one with lowest m/z, thus it is attributed the isotopologue class “M+0”, while B is “M+1”, etc. The real values match the experimental ones. In red, the isotopologue match E-F should exist but it is not found, thus it is thought that F is of class “M+0”. The real values do not match the experimental ones, which need reassignment.

Then the initial isotopologues classification results need to be validated, because some isotopic matches might not be detected and thus the real classes do not match the experimental classes. As the intensity ratio for two consecutive isotopologues reduces as the charge increases, for each type of m/z difference (e.g., “M+0 to M+1”, or “M+1 to M+2”) there is a separate linear relationship between the ratio of two (successive) isotopologues and their m/z value (see Figure 5). In practice, the slope for lines adjusted to each of those types of m/z differences are different, providing a data-driven method to validate the isotopologue matching results, and modify them in case needed (see Figure 6). We defined two methods, Robust Linear Regression Model (RLM) and Linear Support Vector Machine (SVM). RLM is used to model the slope for each type of isotopologue connection (see Figure 7). For additional robustness, the intercept of the regression line should be forced to be zero, since previous research showed that the intercept is close to zero (Hu, 2021). Then a more robust regression line can be calculated for the groups with a small number of matches. After the regression lines for each class are obtained, the standard deviation (STD) of the distance between all matches and the regression line is calculated. If there is only one match in a class, the STD will be set as half of the STD of its previous class. For each match, the weighted distances (distance divided by the STD of distance) to each regression line are calculated and the match is reassigned to the class with the smallest one. In the SVM method, for every two adjacent classes, e.g., M+0 and M+1, a linear SVM model is trained using all m/z and FI ratio pairs of these two classes, and a straight-line boundary that crosses the origin is calculated as the hyperplane to separate these two classes. Then the classification results are calculated by feeding the training data back to the model as test data. Each match is reassigned according to the classification results of SVM.

All isotopologue matches are reclassified into three types, i.e., M+0 to M+1, M+1 to M+2 and M+2 to M+3. The isotopic type for each metabolite feature is assigned accordingly, and features without any isotopologue connections are assigned to M+0.

2.3.2. Thresholds selection for isotopologue matching

As mentioned in the previous section, the initial thresholds for isotopologue matching are large, hence many features are wrongly matched as isotopologues. The accuracy can be increased by reducing the thresholds. However, this will also reduce the number of correct matches. Therefore, an appropriate set of thresholds needs to be decided to reach the balance of correct and incorrect isotopologue matches. An isotopologue match will be considered as an incorrect match if it needs reassignment to a different class, and as a correct match if it does not. The selection of thresholds can be performed using grid-search with the input of three lists for RT, m/z and correlation thresholds respectively. The set of thresholds found using the number of isotopologue connections in each class before and after isotopologue reassignment are later used for adducts as well, since isotopologues and adducts should be found within similar intervals in every dimension. The results are projected into a plot of the number of incorrect matches vs correct matches to measure the matching qualities. To visualise the results, each set of thresholds is encoded into an RGB code to colour the scatter plot, e.g., red increases as m/z decreases; green increases as RT decreases; blue increases with correlation. After obtaining the number of correct and incorrect matches for every set of thresholds in the cubic grid, linear regression is used to model the trend between them, and the furthest set of thresholds below the regression line is selected as the optimal thresholds (see Figure 6).

2.3.3. Intra-dataset adduct matching

An adduct is a product formed by the direct addition of two or more distinct molecules, which contains all atoms of all the components (McNaught and Wilkinson, 1997). They can be formed by chemical entities in contact with the analyte, such as solvents, mobile phase and impurities, thus one metabolite can have different kinds of adducts. We only focus on ten common adducts (Table 1.) in this experiment, as suggested by previous experiments and advice from analytical chemistry experts from the National Phenome Centre. In practice, two features are predicted to be adducts of the same metabolite if they have the same RT, high correlation (the same threshold as for isotopologues) and a specific m/z difference. Each feature yields ten possible neutral masses, which are calculated (using Table 1) as

$$neutral\ mass = \frac{MZ * chargeFactor - massToSubtract + massToAdd}{oligomerIndex} \tag{1}$$

In this workflow, isotopologues were matched first, and thus adducts are matched only between the same type of isotopic features (within isotopologue layers, e.g., M+1, M+2). In each isotopologue layer, if one of the possible neutral mass values for a feature is close to that for another feature, these two features are adducts of the same metabolite and they are matches. Same thresholds of RT, m/z and correlation as in isotopologue matching are used here. In case a feature is matched to more than one feature that will force this feature to be a different adduct, see Supplementary methods for selection.

Table 1. The ten common adducts used in this experiment to construct the metabolite subnetworks. In this example all massToAdd values are zeros, but when including other entities it may have a value, thus it is kept in the table for that generalisation.

ID	chargeFactor	massToSubtract	massToAdd	oligomerIndex
M+H	1	1.007276467	0	1
M+NH4	1	18.03382547	0	1
M+Na	1	22.98922142	0	1
M+CH3OH+H	1	33.03349147	0	1
M+K	1	38.96315942	0	1
M+ACN+H	1	42.03382547	0	1
M+IsoProp+H	1	61.06479147	0	1
M+2H	2	2.014552934	0	1
M+H+NH4	2	19.04110193	0	1
2M+H	1	1.007276467	0	2

2.3.4. Intra-dataset C2H4 matching

Building on the within-dataset matching strategies presented, we explore additional m/z interval matches. Features with similar retention time and a difference of 2 carbons and 4 hydrogens are suspected to be in the same class and have similar properties, hence they can be matched using the m/z difference of 28.031300128 at a small absolute RT difference threshold. Correlation across samples is not guaranteed for these features, so no correlation threshold is set. Once multiple features of the same class have been found, they can be annotated simultaneously by querying databases using an enrichment strategy. Other mass intervals could also be searched for (C2H2, H2, etc), though we have only explored C2H4 to not increase much the number of false positives. To reduce false positives only features of same isotopologue level and adduct type can be matched. A non-linear pattern can be seen in the plot of RT difference vs RT, so we can fit a LOWESS regression to the points (C2H4 matches) in the plot and select the ones within 3*MAD from the regression line to be defined as C2H4 matches.

2.3.5. Matching all features across datasets within thresholds

The inter-dataset difference between any two features is calculated by subtracting the reference feature's value from the target's in RT, m/z and log10 FI. Two features are initially matched if RT difference and m/z difference are both within predefined thresholds. RT difference and m/z difference are changeable parameters adapted to datasets. Multiple matches where one feature is matched to more than one feature in another dataset are allowed at this stage.

2.3.6. Inter-dataset subnetwork matching

Metabolite subnetworks are built by merging all the isotopologue and adduct matchings, where each connected component is a metabolite subnetwork. Figure 3 shows an example of a subnetwork structure. The subnetworks are matched across datasets if there are at least two matched feature pairs between them. A subnetwork in the reference dataset is allowed to be connected to more than one subnetwork in the target dataset if these two subnetworks are not overlapping. These two subnetworks might belong to the same metabolite, but no connection is detected between them. Additionally, those target networks need to be at the same retention time, otherwise, the one with more matched features is the only one selected. If two subnetworks in the target dataset are connected to the same subnetwork in the reference dataset and they overlap, the one with more matched features is selected. In case both have the same number of features, the subnetwork with a smaller averaged m/z difference is selected. After matching subnetworks from reference to target, the results are matched in the other direction, i.e., target to reference, and the common matches are selected.

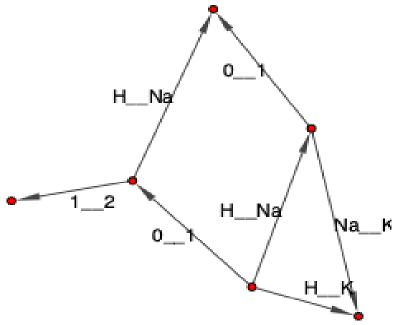


Figure 3. An example of isotopologue-adduct subnetwork. 0_1 and 1_2 are isotopologue matches, and H_Na, N_K, and Na_K are adduct matches.

2.3.7. Select unique matches

A pair of matched features is defined as matched-subnetwork feature pair if both reference and target features appear respectively in an isotopologue-adduct subnetwork with at least two features and these two subnetworks are matched. All matched-subnetwork feature pairs are calculated using the subnetwork matching results, and thus a LOWESS regression of the difference vs the reference value is performed only on matched-subnetwork feature pairs in each dimension separately, to find the inter-dataset shifts (see Figure 9, row 2). A linear interpolation of these LOWESS points allows us to find the expected inter-dataset shift for each point in each dimension. The residuals for all matched feature pairs (not only for the ones in subnetworks) are obtained in each dimension by subtracting the corresponding point in the interpolating curve from each match. The residuals can be standardised by dividing by 5 times MAD of matched-subnetwork feature pairs to allow for its combination into a penalisation score (PS). PS is defined as the weighted sum of squared standardised RT, m/z and log10 FI residuals with weights $W = [w_{RT}, w_{MZ}, w_{FI}]$.

$$PS = w_{RT} * (standRTres)^2 + w_{MZ} * (standMZres)^2 + w_{FI} * (standFIres)^2$$

In many cases FI values are not comparable, thus w_{FI} can be set to 0. In each cluster of multiple matches, the matches with larger penalisation scores are deleted, while all matches not in clusters (only one match for both reference and target feature) are kept.

2.3.8. Delete poor matches

Since large initial thresholds are used, there are some unique matches that are far away from the regression lines in all three domains. They might belong to different metabolites and were wrongly matched together, thus they are defined as poor matches and removed. Larger initial thresholds result in more false-positive matches. To reduce poor matches, a non-linear threshold of RT, m/z and log10 FI is defined as 5 times the MAD of the penalisation scores. All the matches left after deleting poor matches are considered good matches.

2.3.9. Method validation

Compare with known metabolite annotations:

The matching results can be evaluated by comparing with expert annotations where they are available for both reference and target features.

Common-to-Minimum Ratio (CMR) score:

The matching results can be evaluated by calculating a CMR score for each match without using any expert annotations. The overall idea is that a metabolite present in both datasets would be expected to be represented by similar sets of isotopologues and adducts, and these features would be highly correlated with each other in each dataset. If these sets are matched across the datasets, it suggests a successful matching. To calculate the CMR score, first we need to find all highly correlated features separately in reference and target, namely H_1 and H_2 respectively. Highly correlated features are defined as features in the same dataset within small RT distance, and large Pearson correlation, as defined by the isotopologue-adduct thresholds. Subsequently, a good match between any feature in H_1 and H_2 is called a common correlated feature, and the set of all common correlated features for a matched feature is denoted as C. The CMR score for a match is then defined as

$$CMR\ score = \frac{|C|}{\min(|H_1|,|H_2|)+1} \qquad (2)$$

where $|\ X\ |$ is the number of elements in set X. CMR score is in the range $[0, 1)$, where it equals 0 when there is no common correlated feature, and is close to 1 when both $|C|$ and $\min(|H_1|, |H_2|)$ are large. The same-subnetwork features are defined as all features in the same isotopologue-adduct subnetwork. They must be in the same RT and correlation thresholds, therefore, they are a subset of highly correlated features.

3. Results

The Airwave (reference) and MESA (target) datasets (Supplementary Figure 16**Error! Reference source not found.**) mentioned in the previous section were used to demonstrate the entire pipeline of our method. Firstly, the intra-dataset isotopologue-adduct subnetworks were found in each dataset respectively. Then the features were matched across datasets and selected using isotopologue-adduct subnetworks. Finally, we evaluated our matching results.

3.1. Intra-dataset subnetwork construction

3.1.1. isotopologue matching

An initial isotopologue matching with large thresholds (m/z = 0.005 Da, RT = 0.005 min, Pearson correlation = 0.5) was calculated, which ideally captured all possible isotopologue matches in this dataset. Figure 4 shows the distribution of residuals between isotopologues in RT, m/z, and Pearson correlation within the thresholds.

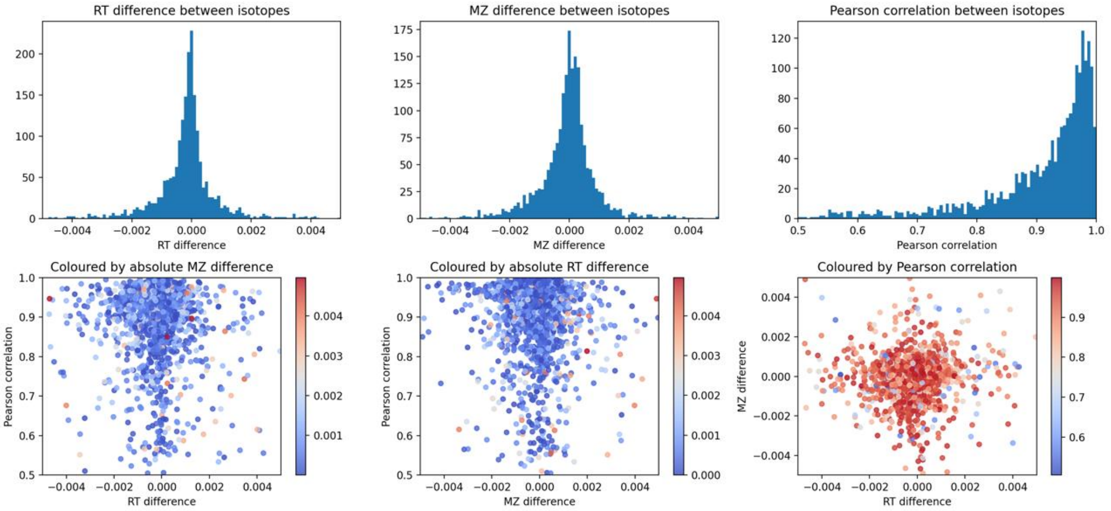


Figure 4. Histograms (top) and scatter plots (bottom) of RT difference, m/z difference, and Pearson correlation of FI between putative isotopologue pairs.

Among all 6327 metabolomic features in Airwave, 1970 isotopologue matches were found, defining 1463 connected components. After all isotopologue matches were obtained, the initial isotopologue connection labels were assigned to each isotopologue match according to the method described in section 2.3.1. The number of isotopologue matches assigned to M+0 to M+1, M+1 to M+2 and M+2 to M+3 was 1463, 421 and 80 respectively. Isotopologue matches assigned to higher-order isotopologues were not considered and deleted at this stage. Figure 5 (left) shows the scatter plot of FI ratio vs m/z coloured by the initial classification results. The FI ratio was calculated by FI of isotopologue $i+1$ divided by FI of isotopologue i, therefore, every FI ratio is between 0 and 1. Obvious patterns can be observed for each isotopologue connection type, which were used to reassign the initial labels that were in the wrong group.

To reassign the isotopologue connection labels, RLM (Figure 5. right) and SVM (Figure 17) were employed. According to the theoretical and practical work mentioned in section 2.3.1, the intercepts were forced to zero in the linear regression for isotopologue connection type classification.

After reassignment using RLM, there were 1443, 455 and 72 isotopologue connections, while using SVM there were 1404, 465 and 101 isotopologue connections for three types of labels respectively. In this experiment, we chose RLM as the primary reassign model, since it generated similar number of isotopologue connections in each class before and after reassignment.

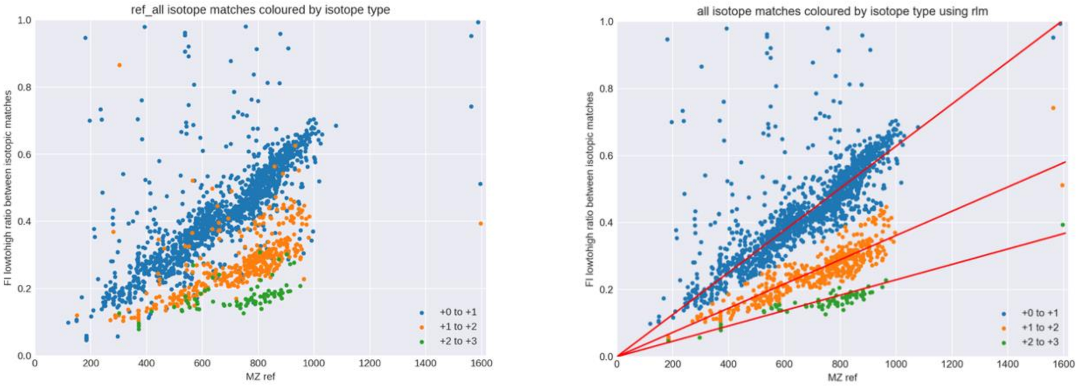


Figure 5. Feature intensity ratio vs m/z scatter plot of (left): all connections within large threshold, (right): reassigned results using RLM. The red lines on the right plot are the regression lines for each class.

3.1.2. Thresholds selection

A grid search of m/z difference between 0.001 and 0.005 (at 0.0005 intervals), RT difference between 0.001 and 0.005 (at 0.0005 intervals), and Pearson correlation between 0.5 and 0.9 (at 0.05 intervals) was performed. The results are shown in Figure 6, where each point is coloured by RGB encoded thresholds. After moving the regression line downwards, an optimal set of thresholds was found with m/z difference=0.002, RT difference=0.0025, and correlation=0.75. There were 1636 isotopologue pairs found in these thresholds, where 1557 matches had consistent isotopologue labels before and after reassignment and 79 matches were reassigned to different classes.

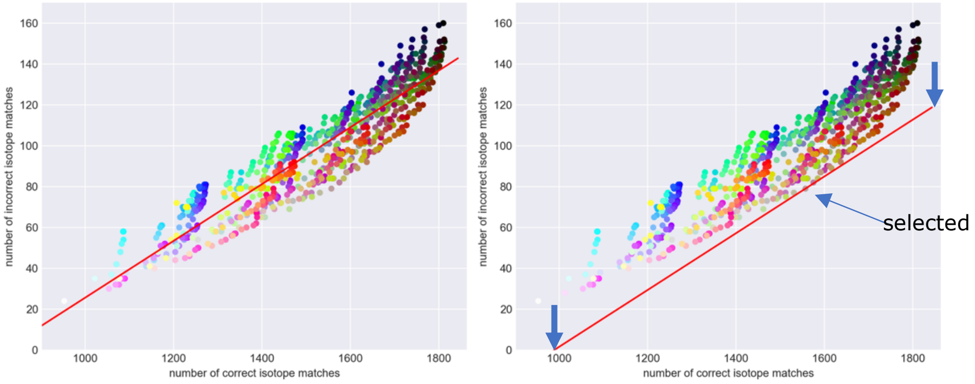


Figure 6. Number of incorrect vs number of correct matches (Left) linear regression results (Right) process of selecting thresholds. The colour used in these plots is RGB code encoded by [RT, m/z, correlation] rescaled from [[0.005, 0.001], [0.005, 0.001], [0.5, 0.9]] to [[0, 1], [0, 1], [0,1]].

Figure 7 shows the FI ratio vs m/z plot before and after isotopologue connection reassignment within the selected optimal thresholds. On the left plot, there were 1272, 318, 46 isotopologue matches for each type respectively, which changes to 1235, 358, 43 on the right plot. There were 2645 metabolomic features without any isotopologue matches, thus assigned to type M+0.

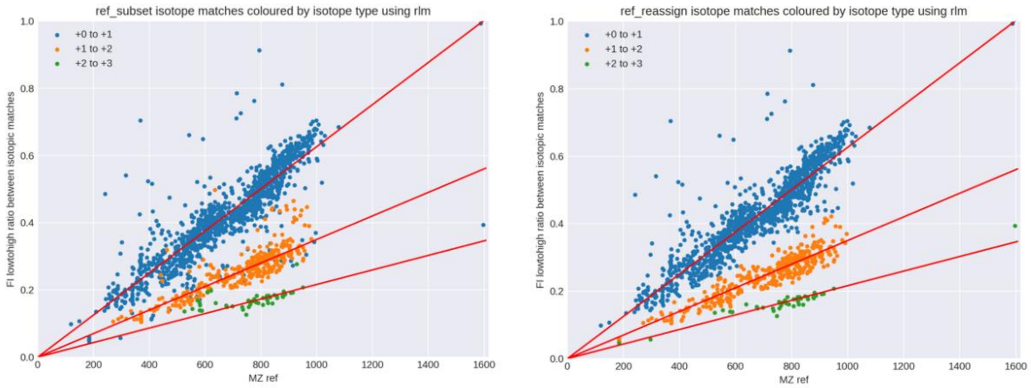
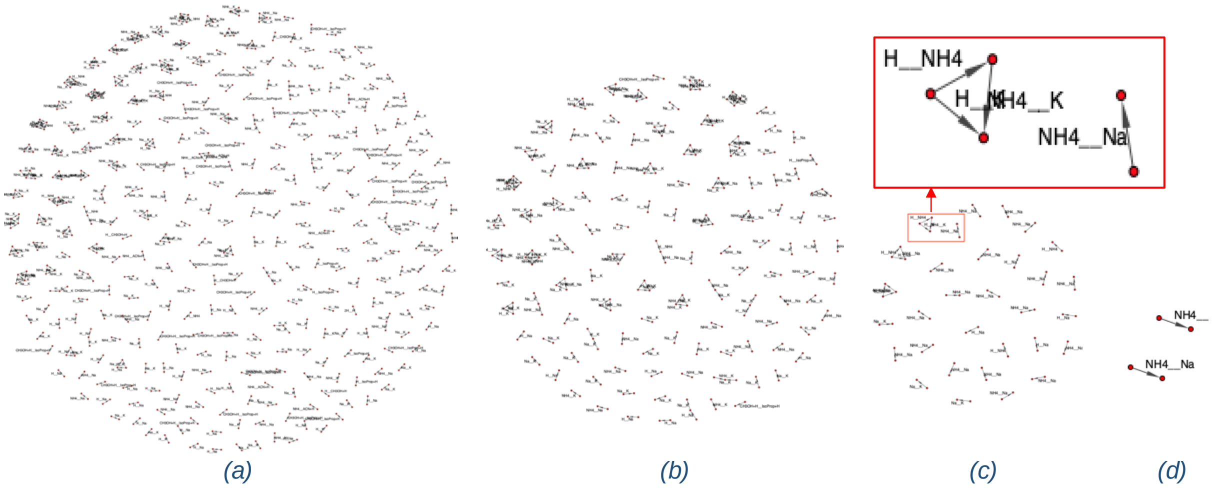


Figure 7. FI ratio vs m/z before (left) and after (right) reclassification using RLM with the optimal set of thresholds.

3.1.3. Adduct matching

After obtaining the isotopologue type label for each metabolomic feature, adduct matching was performed within each layer of isotopologue separately. Only 10 commonly detected adducts were used in the adduct matching process. 656 adduct matches were initially found within the thresholds defined during isotopologue matching. After 34 multiple matches were deleted, 428, 189, 37 and 2 adduct matches were left in each isotopologue layer respectively. Figure 8 shows the adduct subnetworks within each isotopologue layer, where the nodes are metabolomic features, and the edges are adduct connections. There are 290, 110, 31 and 2 connected components in each layer. The three most frequent adduct connections (between two adducts of the same metabolite, e.g., Na and K adducts) are Na_K (179), H_Na (145), and NH4_Na (143).



(a) (b) (c) (d)

Figure 8. Adduct matching results within isotopologue layers M+0 (a), M+1 (b), M+2 (c), and M+3 (d).

Finally, the subnetworks were constructed in each dataset by merging all isotopologue and adduct matches, resulting in 1193 subnetworks with at least two features. We applied the same procedure for the target dataset obtaining 1322 subnetworks.

3.2. Feature matching across datasets

3.2.1. Match all features within thresholds

All possible feature matching pairs between reference and target datasets were calculated within large thresholds (RT = 0.1 min, m/z = 0.015 Da, FI ref > FI tar) to capture the trends of differences between them. The inter-dataset differences can be clearly observed in all three dimensions on the 1st row of Figure 9. The isotopologue type and adduct type were not considered for initial matching. There are 6327 and 10910 features in the reference and target dataset respectively, from which we obtained 2927 initial matches. 2908 features in reference were matched to 2797 features in the target dataset. In the reference dataset, 19 matches were double matches, and 2889 were single matched. In the target dataset, 2 features had 3 matches, 126 features had double matches, and 2669 were single matches.

3.2.2. Subnetwork matching

467 isotopologue-adduct subnetworks in the reference were matched to 435 in the reference. There had to be at least two matched feature pairs between each pair of subnetworks. Among 2927 initial feature matches, 1189 matched-subnetwork feature pairs were obtained by using the subnetwork matching results.

3.2.3. Select unique matches and delete poor matches

A LOWESS regression was applied using only the matched-subnetwork feature pairs. In 2nd row of Figure 9, green dots are the matched-subnetwork feature pairs, and the red line is the LOWESS regression line. A fraction of 0.1 (10% of the points) was used in LOWESS regression for all domains. In each dimension, the residuals were obtained for each match (Figure 9, 3rd row), and standardised (Figure 9, 4th row) dividing by 5 times the MAD in the respective dimension. Penalisation scores for each match were calculated as the weighted sum using a weight $W_{RT,MZ,FI} = [1,1,0]$ of the squared standardised difference so that RT and m/z influenced the score equally and FI was ignored. The visualisation of the penalisation score is shown in Figure 9. 5th row.

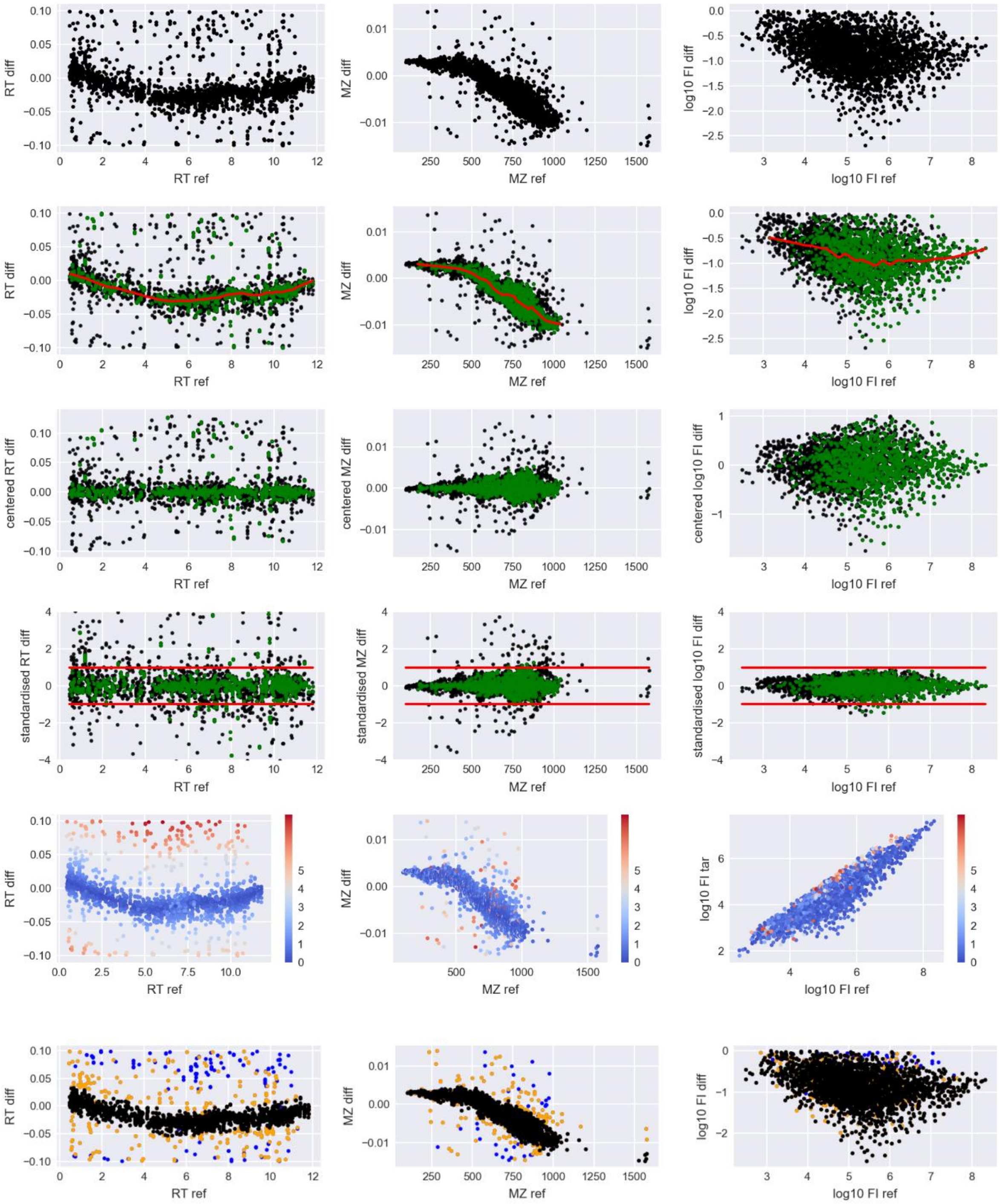


Figure 9. Workflow of matching Airwave and MESA. 1st row: inter-dataset difference for matched feature pairs in RT, m/z and log10 FI. Each black dot is an initial feature match. 2nd row: green dots are matched-subnetwork feature pairs, and the red line is the interpolated LOWESS regression using only green dots. 3rd row: the residuals from each point to the red line. 4th row: standardised residuals by dividing by 5 times MAD respectively. 5th row: the same plot coloured by penalisation score with weight W = [1, 1, 0]. 6th row: final matching results with 5 times penalisation score as the threshold where blue points are deleted multiple matches, yellow points are poor matches, and black points are good matches.

After selecting the matches with the smallest penalisations score within each cluster of multiple matches, 2778 single matches were found from 2927 initial matches (95.0%). A non-linear threshold was defined as 5*MAD of the penalisation score to remove poor matches, which led to

2465 good matches as the final result. These results are shown in Figure 9 6th row, where blue points are deleted multiple matches, yellow points are poor matches, and black points are good matches.

3.3. Method validation

3.3.1. Compare with known metabolite annotations

The feature matching was evaluated using dataset-specific expert metabolite annotations, its results are reported in Figure 10 and a summary of the number of annotations and matches at each stage is shown in Table 2. There were 567 known common annotations available for both reference and target features. After defining the initial thresholds, 16 annotations were outside of the initial thresholds and therefore not matched. After deleting multiple matches, all the remaining 551 annotations were correctly selected among their multiple matches. After tightening the thresholds, 24 annotations were classified as poor matches and incorrectly deleted. Among 2465 good matches, 527/567 (92.9%) annotated matches were found correctly, and there is only one wrong match which had inconsistent annotations. 289/313 (92.3%) poor matches were deleted correctly.

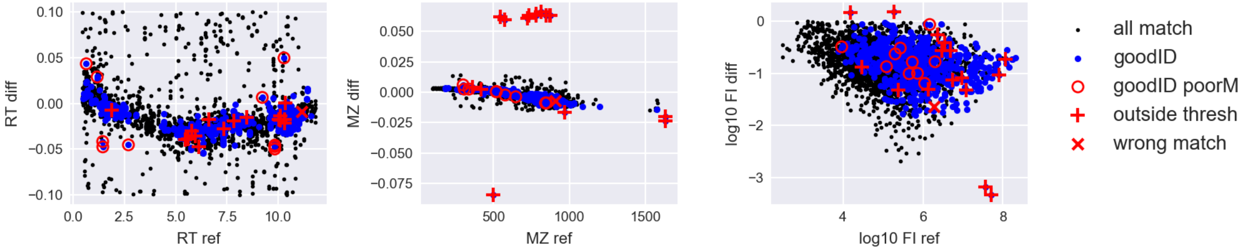


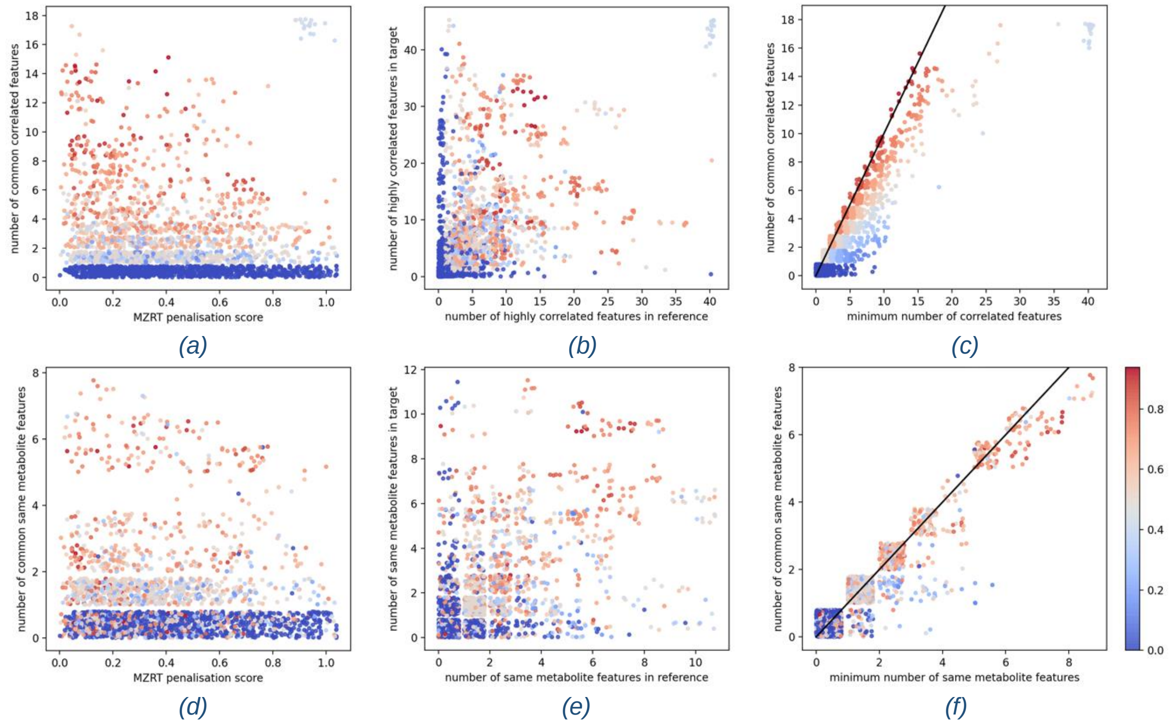
Figure 10. Feature matching results shown in RT, m/z and log10 FI domains. Black dots are all matches within the initial thresholds, blue dots are all the matches with identical annotations, red circles are matches with identical annotations that are incorrectly considered as poor matches, + symbols are matches with identical annotations that are outside of initial thresholds, and x symbol is an incorrect match.

Table 2. Summary of number of annotations and matches at each stage.

	Annotations	Matches
All	567	
Outside initial thresholds	16	
Within initial thresholds	551	2927
Deleted multiple matches	0	149
Single matches	551	2778
Deleted poor match	24	313
Good matches	527	2465
Wrong matches		1

3.3.2. CMR score

Within each dataset, highly correlated (Pearson correlation > 0.75) features within absolute RT difference (< 0.0025 min) were detected for each reference and target feature respectively (H_1 and H_2), and the number of common correlated features (C) was determined. CMR scores were calculated following Equation 2. In Figure 11, all plots are coloured by the same CMR score, but they have different axes. The axes of plots (a, b, c) consider all features, and (d, e, f) consider only the ones in isotopologue-adduct subnetworks. (a) and (d) show that there tend to be more common features for matches with low penalisation scores. (d) also shows good agreement between penalisation score and CMR score, where lower penalisation scores correspond to higher CMR scores. (e) shows more consistency than (b) between two datasets. (c) and (f) show that most of the features in the smaller H are matched. (d), (e) and (f) show that CMR score is a stable evaluation method.



(a) (b) (c)

(d) (e) (f)

Figure 11. Evaluation plots are coloured by CMR score. The x-axis is jittered in (b), (c), (e) and (f), and the y-axis is jittered in all plots. Top plots consider all features, while the bottom plots consider only the ones in isotopologues-adduct subnetworks.

3.4 C2H4 matching

An RT threshold of 1 minute and m/z threshold same as isotopologue-adduct threshold is used in this experiment for C2H4 matching. Figure 12 shows the C2H4 matching results in RT difference vs RT plots for the Airwave dataset. The left plot is C2H4 matches (3615) using all features in the reference datasets, and the right plot are C2H4 matches (865) using only features that are matched across datasets. The red line is the linear interpolator of LOWESS regression of all matches in each plot respectively. There were 3027 and 638 matches within 3*MAD of all points respectively, which are coloured in blue.

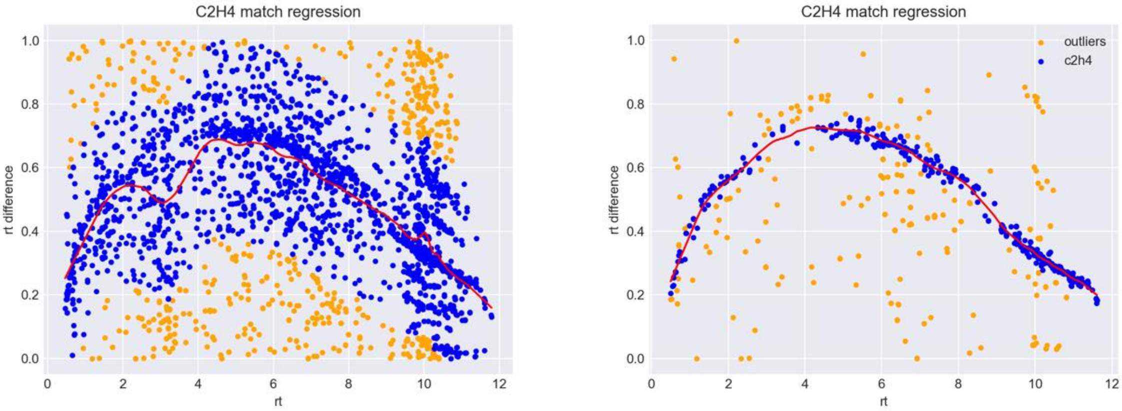


*Figure 12. RT difference vs RT for C2H4 matches with the LOWESS regression in the Airwave dataset. The left plot is C2H4 matches using all features in the reference datasets, the right plot is C2H4 matches using only features that are matched across datasets. The red line is the linear interpolator of LOWESS regression. Blue dots are within 3*MAD of all points, and yellow dots are outside.*

3.5. Matching of other datasets

MESA phase 1 and MESA phase 2

Here we evaluated the workflow on a very challenging example, matching of datasets acquired with very different chromatographic gradients. Due to the different chromatographic gradients employed, most of the features in MESA 1 have RT between 0 and 8, while in MESA 2 between 0 and 12 (see Supplementary Figure 18).

The same procedure was performed for isotopologue and adduct matching in MESA 1 as described in previous sections. RT threshold was set a different value of [-1, 5] minutes when matching across datasets with MESA 2, and m/z thresholds were the same [0.015, 0.015] Da. The matching results are shown in Figure 13, where we used MESA 1 as reference and MESA 2 as target by random choice. There are 3241 and 10910 features in reference and target datasets respectively, from which we obtained 2557 initial matches. 1858 features in reference were matched to 2456 features in the target dataset. In the reference dataset, 519 matches were in clusters of multiple matches, and 1339 were only in one match. In the target dataset, 101 features were in clusters of multiple matches, and 2355 were only in one match. Among 2557 initial matches, 515 subnetwork-matched feature pairs were calculated by using the subnetwork matching results. After deleting multiple matches and poor matches, 1837 single matches were left (71.8%) and 1466 good matches were kept (57.3%). The evaluation of the matching results using CMR score can be seen in Supplementary Figure 19.

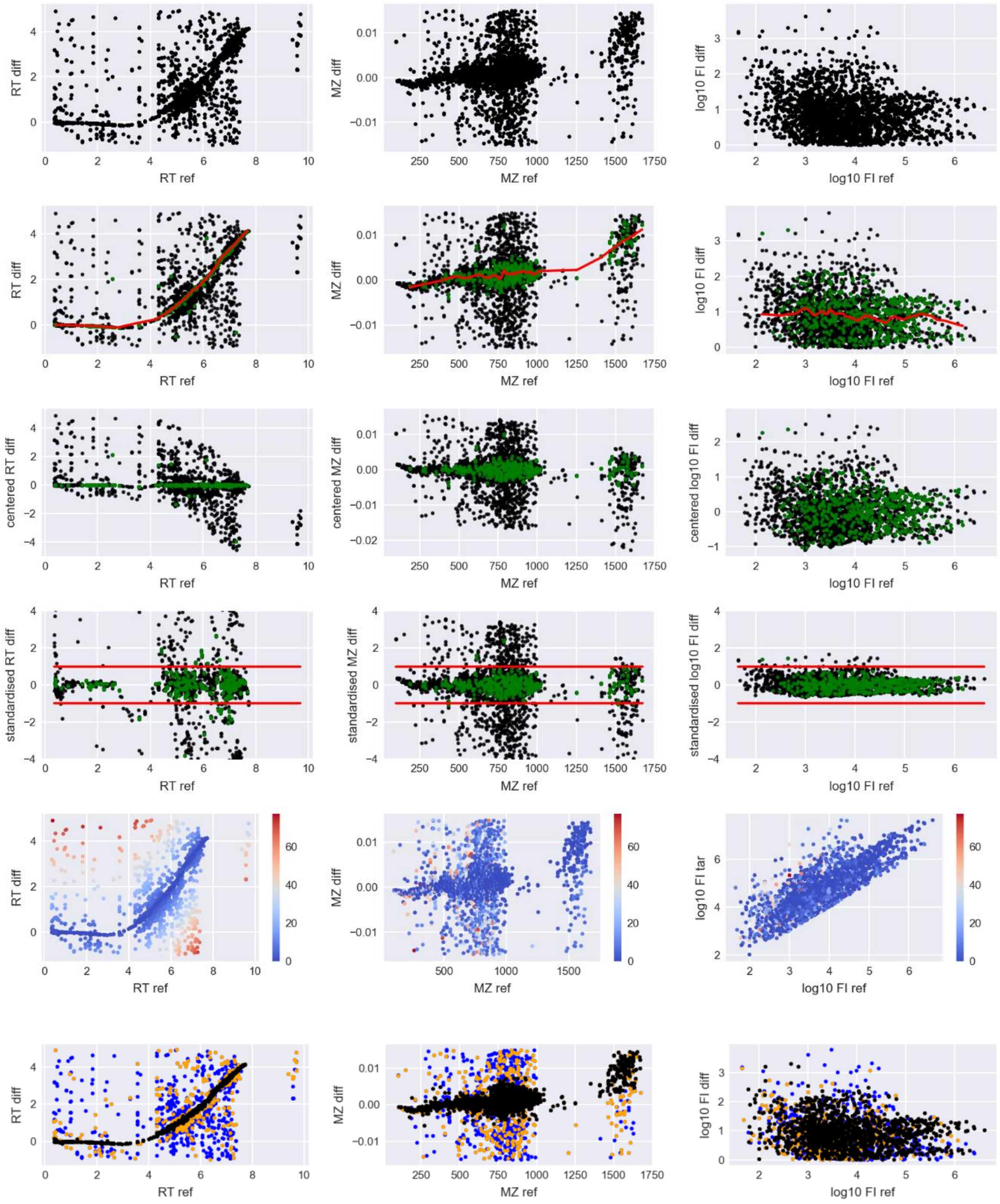


Figure 13. Workflow of matching MESA phase 1 and MESA phase 2.

4. Discussion

The study of LCMS metabolomics has been constrained by the difficulty of merging datasets acquired under different conditions. At this moment, most of the related software does sample-to-sample feature matching but not dataset-to-dataset feature matching. The datasets we used for inter-dataset feature matching are obtained by feature matching across samples, where the median value of RT, m/z and FI across samples are used as the corresponding feature values of the dataset. The metabolomics features could simply be matched across datasets using metabolite identifications. However, it is very time-consuming to identify metabolites, so it is impossible to match everything by using labels.

In this paper, we constructed intra-dataset network structures to improve the process of feature matching across datasets. The simplest of the dataset-to-dataset matching strategies is to set absolute thresholds for RT and select the one with the closest m/z value within the threshold, which is a basic version of our work. In (Habra et al., 2021) features at similar m/z and retention times are first binned, then “anchor” matches are found one-by-one using the features of higher intensity only. In opposition to ours, this method aligns only the RT dimension and can only work if the FI of matches are correlated.

Comparing to our baseline model by Pinto et al. (2021) using one-to-one feature matching, an intra-dataset network structure improves the robustness for finding inter-dataset shifts by only using high-quality feature matches. This reduces the number of false-positive matches, however at the same time requires the existence of isotopologues and adducts, in both datasets.

Other fields such as proteomics or transcriptomics do not find the same annotation problems as untargeted metabolomics. In untargeted proteomics, proteins have known sequences of amino acids that can be compared to databases and their IDs found. In transcriptomics, RNA sequences are broken into nucleic acid chains, which can also be compared to databases.

4.1. Correlation calculation for isotopologue and adduct matching

Our strategy of using different percentiles to calculate the correlations is important because of the lower detection limit and detector saturation, as well as random outliers. The correlations are calculated using four different percentiles and the highest one is chosen among them. These correlations should all be the same for one feature. However, because samples are too diluted to be well measured on the low-intensity side, and too concentrated to be well measured on the high-intensity side, this has a different effect on adducts and especially isotopologues (e.g., isotopologues M+3 have much lower intensity than M+0 and may not be well measured in the lower ranges).

4.2. Isotopologue re-labelling method

The low-to-high FI ratio of isotopologues should be between 0 and 1, but there could be wrongly matched isotopologues also with FI ratio between 0 and 1 or even above 1. In Figure 7, there are some blue points far above the blue cluster (M+0 to M+1), and they might be false-positive isotopologue matches. Therefore, in the future, a boundary should be defined to reduce false positives. RLM and SVM were implemented for isotopologue re-labelling. When matching Airwave and MESA datasets, the reassigned results by RLM have a smaller change in the number of isotopologue connections in each class before and after reassignment than SVM. The number of +2 to +3 matches increased from 80 to 101 when using SVM and decreased to 72 when using RLM. If one more match in +2 to +3 class is found, there must be a corresponding new +0 to +1 connection and a +1 to +2 connection. However, the number of connections of +0 to +1 reduced more using SVM, which implies more potential inconsistency after reassignment. Additionally, some +0 to +1 connections were initially classified as +1 to +2 due to the outliers in +0 to +1 class, which skewed the reclassification results of SVM to have more higher-order isotopologue matches (Figure 5 and Figure 17). However, this impacts RLM less since it uses a robust model to reduce the effect of outliers. Therefore, RLM was selected as the reassign model.

4.3. Thresholds selection method

Larger initial thresholds for isotopologue matching result in more false-positive matches, whereas by reducing the thresholds more correct matches are deleted. Our thresholds selection method suggests a locally optimal set of thresholds. A trend is found between the number of incorrect matches and correct matches, which is in effect weighing the costs of incorrect matches and correct matches in a specific way according to the dataset. In our experience, near-optimal thresholds give similar possible good solutions. Nevertheless, further work might be needed to decide how to select the globally optimal thresholds.

4.4. Validation approaches and their results

In our main example there were 16 annotations ae outside initial thresholds, and far away from the trend in the m/z dimension. Penalisation scores are calculated using a weight of [1, 1, 0] from each dimension. This gives equal weights to RT and m/z, and no effect from FI as the datasets were of plasma and serum, and thus FI was not necessarily comparable. The default threshold for poor matches is 5 times MAD of penalisation scores. A smaller threshold results in deleting more feature matches, including both true negative matches (poor matches with different identifications) and false negative matches (poor matches with identical identifications). When deleting multiple matches, all the annotations were kept, which means all multiple matches were selected correctly for those with known labels. In this case, 24/551 (4.36%) annotations and 313/2778 (11.27%) feature matches were deleted, and false positives were removed at a much higher rate. Therefore, deleting poor matches is an effective and necessary step to reduce false positives.

In Figure 11, the first row shows the pattern of all highly correlated features coloured by CMR score, and the second row shows the pattern of only high confidence isotopologues and adducts coloured by the same CMR score as in the first row. Hence, the second row is the validation of the first row, because the matches presented are more robust and it still shows the same trends. The CMR scores show good agreement with both penalisation score and the number of common same-metabolite features in plot (d). Plots (b, e) show that there are more isotopologues and adducts in the target dataset, which is natural because there are much more features in the target dataset.

4.5. C2H4 matching

The matching of same-class metabolomic features was attempted, by means of C2H4 differences. The challenge in implementing the method is the large number of false positives and the lack of well-defined trends (in Figure 12 left). Contrarily to the isotopologues, at this point, there is no numerical way to validate if the C2H4 matches are correct or not. The strategy was successfully implemented in the software, though it is not used for inter-dataset matching since there were too many false positives. It may be useful in a future iteration of the software for intra-dataset matching or to provide putative annotations from databases for groups of same-class features using enrichment strategies. We suggest that this annotation strategy should be applied for example after inter-dataset matching, which serves as a filter to minimise false-positive C2H4 matches later.

4.6. Matching of MESA 1 and MESA 2

In Figure 18, we show retention times from 0 to 15 minutes to make the visualisation comparable to each other. However, the usable part of the chromatogram does not start at 0 minutes, because there are many molecules that are not retained by the column. Similarly, after 8 minutes the detected features are not meaningful anymore. In practice, the features around 0 and after 8 minutes could be deleted before any matching since they are not relevant.

The difficulty of matching two datasets depends on the RT, m/z and FI dissimilarity. Matching MESA 1 and MESA 2 is very challenging because they were acquired with very different chromatographic gradients and flow rates. Due to this, the range of RT of MESA 1 is much smaller than the RT range of MESA 2, which creates a strong non-linear relationship across datasets in RT and force the use of large RT thresholds. There is a much larger difference in RT between MESA 1 and 2 than between Airwave and MESA 2 since these used more similar conditions. A larger RT threshold naturally increases the number of false positives. Besides, features are more concentrated from 4 min to 6 min in MESA 1, which also results in more false-positive matches.

4.7. Future Work

4.7.1. Matching multiple datasets

Our current method only match features across two datasets, and further work could explore adapting our method to match across multiple datasets. The intra-dataset subnetworks would be constructed as usual. Then initial feature matches could be found between every two datasets. Features matched across more than a predefined number of datasets (e.g., 3) would be kept, and others deleted. This intermediate step should help reduce the number of multiple match clusters. The algorithm then would then proceed similarly to ours.

4.7.2. Using Bayesian statistics to model the inter-dataset shift

The inter-dataset shifts are modelled using LOWESS in RT, m/z and FI separately in our method. Bayesian statistics could be used as an even more robust strategy to model the inter-dataset shift with a prior that is calculated using only matched-subnetwork feature pairs and the posterior distribution that is calculated using all feature matches within thresholds. This method would use the information from three dimensions (RT, m/z, FI) simultaneously instead of modelling the shift in each dimension separately.

4.7.3. Isotopologue reassignment

The current isotopologue reassignment methods are RLM and SVM, which both set a hard straight-line boundary between every two adjacent classes. We suggest that in the future special attention is devoted to the reassignment of matches close to these boundaries. In Figure 5 (left), some orange points are closer to the cluster of blue points. They were reclassified into the blue class by using either RLM or SVM. However, these isotopologue connections are in the M+1 to M+2 class, which means that their corresponding M+0 to M+1 connections were detected. This evidence is stronger than using the gradient when the connections are near the boundary. For the blue points that are in the orange cluster, they might belong to M+0 to M+1 class, or M+1 to M+2 class if their corresponding M+0 to M+1 connections are not detected. For M+2 to M+3 matches, their feature intensity values are small, so they are measured with more errors. Therefore, a more accurate isotopologue reassign method is stringent.

4.7.4. C2H4 matches for identification

After matching features across datasets, C2H4 connections could be used for metabolite annotation. A compound with a specific m/z and RT might correspond to many different annotations in databases. A sequence of the compounds that increase with C2H4 suggests it is the same class of compounds, which could be used to reduce the number of possible annotations when comparing to databases. Additionally, the C2H4 connected features may have already known isotopologue and adduct type, limiting even more the annotation possibilities.

4.7.5. Iso-directional method

At this moment, our method gives us slightly different results for matching two datasets in different directions, i.e., from reference to target, and from target to reference. The reason is that during isotopologue reassignment and finding inter-dataset shifts, the results depend on the value of the reference feature. We could explore the possibilities to define an iso-directional method, from which we will obtain the same matching results no matter which dataset is used as the reference. One possible strategy is to use the same strategy also in the target – reference direction, then select the common findings, as it is used for isotopologue-adduct subnetworks inter-dataset matching.

5. Conclusion

In this paper, we proposed a method to find the same features across two datasets in LC-MS untargeted metabolomics studies. We constructed intra-dataset networks and used them to find the RT, m/z and FI shifts between two datasets, which allowed the creation of a penalisation score for each match. We also designed an evaluation method without the use of annotations, which showed that the matches found were of high qualities. The method was implemented as a Python package, publicly available.

References

ÅBERG, K. M., ALM, E. & TORGRIP, R. J. 2009. The correspondence problem for metabonomics datasets. *Analytical and bioanalytical chemistry,* 394**,** 151-162.

DETTMER, K., ARONOV, P. A. & HAMMOCK, B. D. 2007. Mass spectrometry‐based metabolomics. *Mass spectrometry reviews,* 26**,** 51-78.

DONA, A. C., JIMÉNEZ, B., SCHÄFER, H., HUMPFER, E., SPRAUL, M., LEWIS, M. R., PEARCE, J. T., HOLMES, E., LINDON, J. C. & NICHOLSON, J. K. 2014. Precision high-throughput proton NMR spectroscopy of human urine, serum, and plasma for large-scale metabolic phenotyping. *Analytical chemistry,* 86**,** 9887-9894.

DUNN, W. B., BROADHURST, D., BEGLEY, P., ZELENA, E., FRANCIS-MCINTYRE, S., ANDERSON, N., BROWN, M., KNOWLES, J. D., HALSALL, A. & HASELDEN, J. N. 2011. Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. *Nature protocols,* 6**,** 1060-1083.

FAHY, E., SUD, M., COTTER, D. & SUBRAMANIAM, S. 2007. LIPID MAPS online tools for lipid research. *Nucleic acids research,* 35**,** W606-W612.

HABRA, H., KACHMAN, M., BULLOCK, K., CLISH, C., EVANS, C. R. & KARNOVSKY, A. 2021. metabCombiner: Paired Untargeted LC-HRMS Metabolomics Feature Matching and Concatenation of Disparately Acquired Data Sets. *Analytical Chemistry,* 93**,** 5028-5036.

HU, M. 2021. *Molecular networking for the fusion of multiple untargeted LC-MS metabolomic datasets.* MRes Biomedical Research, Imperial College London.

IZZI‐ENGBEAYA, C., COMNINOS, A. N., CLARKE, S. A., JOMARD, A., YANG, L., JONES, S., ABBARA, A., NARAYANASWAMY, S., ENG, P. C. & PAPADOPOULOU, D. 2018. The effects of kisspeptin on β‐cell function, serum metabolites and appetite in humans. *Diabetes, Obesity and Metabolism,* 20**,** 2800-2810.

KUHL, C., TAUTENHAHN, R., BOTTCHER, C., LARSON, T. R. & NEUMANN, S. 2012. CAMERA: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. *Analytical chemistry,* 84**,** 283-289.

LEWIS, M. R., PEARCE, J. T., SPAGOU, K., GREEN, M., DONA, A. C., YUEN, A. H., DAVID, M., BERRY, D. J., CHAPPELL, K. & HORNEFFER-VAN DER SLUIS, V. 2016. Development and application of ultra-performance liquid chromatography-TOF MS for precision large scale urinary metabolic phenotyping. *Analytical Chemistry,* 88**,** 9004-9013.

MCNAUGHT, A. D. & WILKINSON, A. 1997. *Compendium of chemical terminology*, Blackwell Science Oxford.

PINTO, R. C. K., IBRAHIM; LEWIS, MATTHEW R.; HÄLLQVIST & JENNY; KALUARACHCHI, M. G., GONÇALO; CHEKMENEVA, ELENA; GRIFFIN, JULIAN; DEHGHAN, ABBAS; ELLIOTT, PAUL; TZOULAKI,IOANNA; HERRINGTON, DAVID; EBBELS, TIMOTHY 2021. Finding correspondence between metabolomic features in untargeted liquid chromatography - mass spectrometry datasets. *Manuscript in preparation. .*

PLUSKAL, T., CASTILLO, S., VILLAR-BRIONES, A. & OREŠIČ, M. 2010. MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC bioinformatics,* 11**,** 1-11.

SCHRIMPE-RUTLEDGE, A. C., CODREANU, S. G., SHERROD, S. D. & MCLEAN, J. A. 2016. Untargeted metabolomics strategies—challenges and emerging directions. *Journal of the American Society for Mass Spectrometry,* 27**,** 1897-1905.

SMITH, C. A., WANT, E. J., O'MAILLE, G., ABAGYAN, R. & SIUZDAK, G. 2006. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Analytical chemistry,* 78**,** 779-787.

TSUGAWA, H., CAJKA, T., KIND, T., MA, Y., HIGGINS, B., IKEDA, K., KANAZAWA, M., VANDERGHEYNST, J., FIEHN, O. & ARITA, M. 2015. MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis. *Nature methods,* 12**,** 523-526.

WANG, M., CARVER, J. J., PHELAN, V. V., SANCHEZ, L. M., GARG, N., PENG, Y., NGUYEN, D. D., WATROUS, J., KAPONO, C. A. & LUZZATTO-KNAAN, T. 2016. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nature biotechnology,* 34**,** 828-837.

WISHART, D. S. 2008a. Applications of metabolomics in drug discovery and development. *Drugs R D,* 9**,** 307-22.

WISHART, D. S. 2008b. Quantitative metabolomics using NMR. *TrAC trends in analytical chemistry,* 27**,** 228-237.

XIA, J., BROADHURST, D. I., WILSON, M. & WISHART, D. S. 2013. Translational biomarker discovery in clinical metabolomics: an introductory tutorial. *Metabolomics,* 9**,** 280-299.

ZHOU, B., XIAO, J. F., TULI, L. & RESSOM, H. W. 2012. LC-MS-based metabolomics. *Molecular BioSystems,* 8**,** 470-481.

Appendix

Supplementary methods

Isotopologue multiple matching

When matching isotopologues, there will be some features that should not have been matched together because of the use of large thresholds. The types of initial isotopologues matches are decided solely by the relative m/z value in their connected component. Therefore, the wrongly matched pairs will potentially disturb the isotopologue match classification. We defined m/z difference to be the only factor when selecting the valid match from multiple isotopic matches. An example is shown in Figure 14.

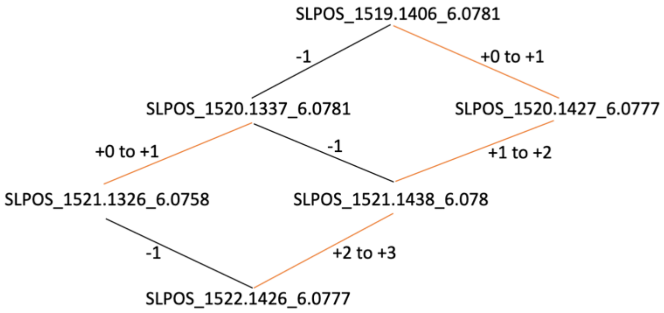


Figure 14. Example of selection from multiple isotopic matches in MESA phase 2. Black connections will be deleted from initial matches and marked as -1. Orange connections will be selected as valid isotopologue matches. Two groups of isotopologues are detected in this initial isotopologue group.

Adduct multiple matching

During adduct matching, one feature can be connected to more than one feature if all the connections indicate this feature to be the same adduct. If a feature has multiple connections which forces it to be different adducts, then the adduct connections are selected by the occurrences of references type. If two adduct type have the same number of connections, then the adduct type is decided by the occurrences of adduct match type. In the example in Figure 15, the adduct match Na_K was deleted first. If the number of H_NH4 plus H_Na in the entire dataset is larger than the number of NH4_Na plus NH4_K, then the reference feature is selected to be of adduct type +H, and adduct matches H_NH4 and H_Na is kept. Otherwise, the reference is chosen to be +NH4, and adduct matches NH4_Na and NH4_K are kept.

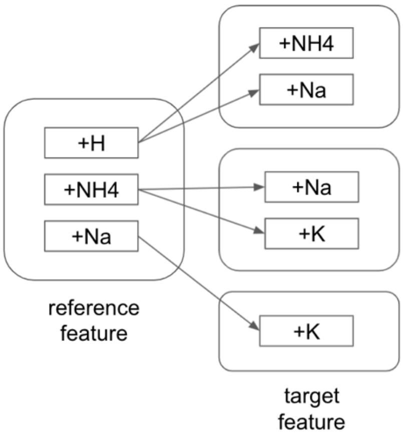
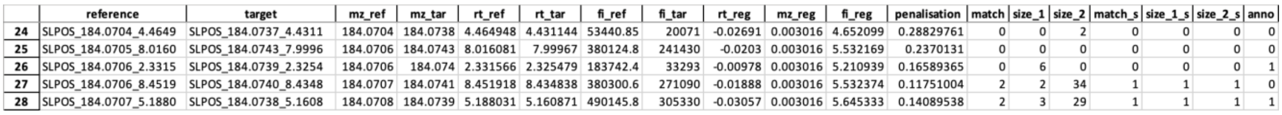


Figure 15. An example to show the process of selecting from multiple adduct matches.

Supplementary figures and tables

Table 3 shows five example of feature matches where each row is a match. The penalisation scores are used to delete multiple matches and reduce poor matches. Columns 13-15 were used to calculate the CMR scores and plot the first row of figure 11, and the columns 16-18 were used to plot the second row.

Table 3. Example of output where each row is a feature match across datasets. The first two columns are feature names for the reference and target feature. The 3-8 columns are m/z, RT and FI values for the reference feature and the target feature respectively. The 9-11 columns are interpolated LOWESS regression value in RT, m/z and FI. The 12 column is the penalisation score. The 13 column is the number of common correlated features. The 14-15 columns are the number of highly correlated features in the reference dataset and the target dataset. The 16 column is the number of common same-network features. The 17-18 columns are the number of same-network features in the reference dataset and the target dataset.



	reference	target	mz_ref	mz_tar	rt_ref	rt_tar	fi_ref	fi_tar	rt_reg	mz_reg	fi_reg	penalisation	match	size_1	size_2	match_s	size_1_s	size_2_s	anno
24	SLPOS_184.0704_4.4649	SLPOS_184.0737_4.4311	184.0704	184.0738	4.464948	4.431144	53440.85	20071	-0.02691	0.003016	4.652099	0.28829761	0	0	2	0	0	0	0
25	SLPOS_184.0705_8.0160	SLPOS_184.0743_7.9996	184.0706	184.0743	8.016081	7.99967	380124.8	241430	-0.0203	0.003016	5.532169	0.2370131	0	0	0	0	0	0	0
26	SLPOS_184.0706_2.3315	SLPOS_184.0739_2.3254	184.0706	184.074	2.331566	2.325479	183742.4	33293	-0.00978	0.003016	5.210939	0.16589365	0	6	0	0	0	0	1
27	SLPOS_184.0706_8.4519	SLPOS_184.0740_8.4348	184.0707	184.0741	8.451918	8.434838	380300.6	271090	-0.01888	0.003016	5.532374	0.11751004	2	2	34	1	1	1	0
28	SLPOS_184.0707_5.1880	SLPOS_184.0738_5.1608	184.0708	184.0739	5.188031	5.160871	490145.8	305330	-0.03057	0.003016	5.645333	0.14089538	2	3	29	1	1	1	1

Figure 16 shows the Airwave (reference) and MESA (target) datasets for our main example.

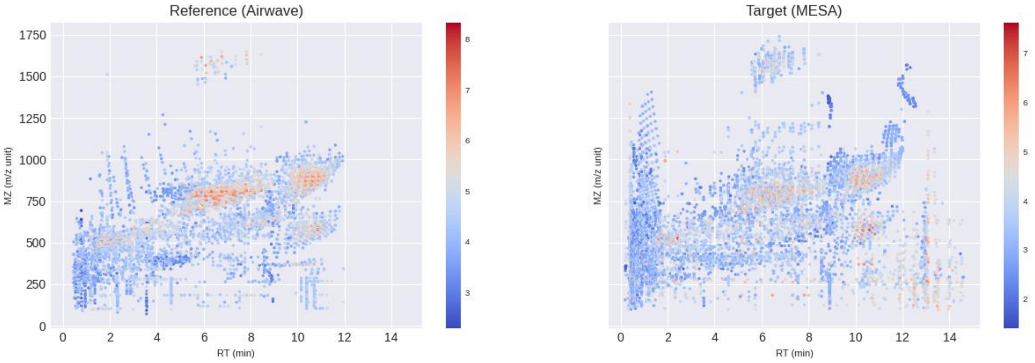


Figure 16. m/z vs RT plots of Airwave reference dataset (left) and MESA target dataset (right) coloured by log10 FI.

Figure 17 shows the reassignment results using SVM for all isotopologue matches.

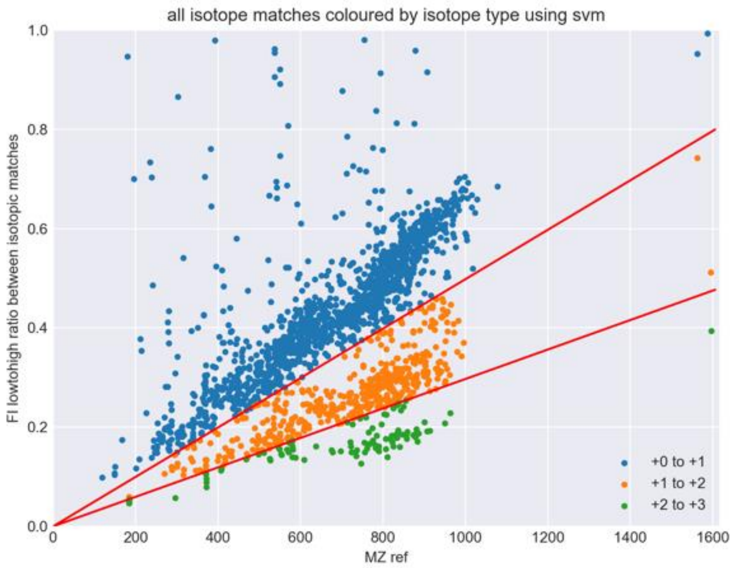


Figure 17. Reassigned results using SVM, and the red lines are the hyperplanes.

Figure 18 shows the MESA 1 (reference) and MESA 2 (target) datasets for the second example.

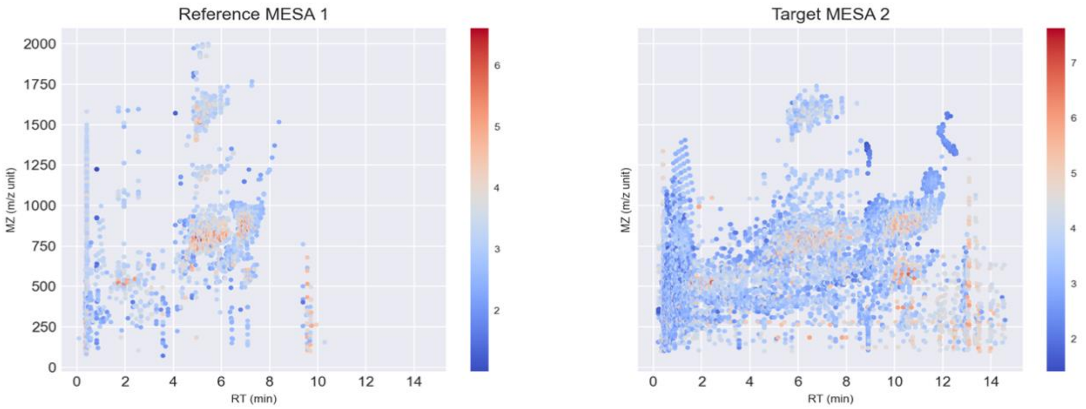


Figure 18. m/z vs RT plots of MESA phase 1 reference dataset (left) and MESA phase 2 target dataset (right) coloured by log10 FI.

Figure 19 shows the evaluation of the matching results of MESA 1 and MESA 2.

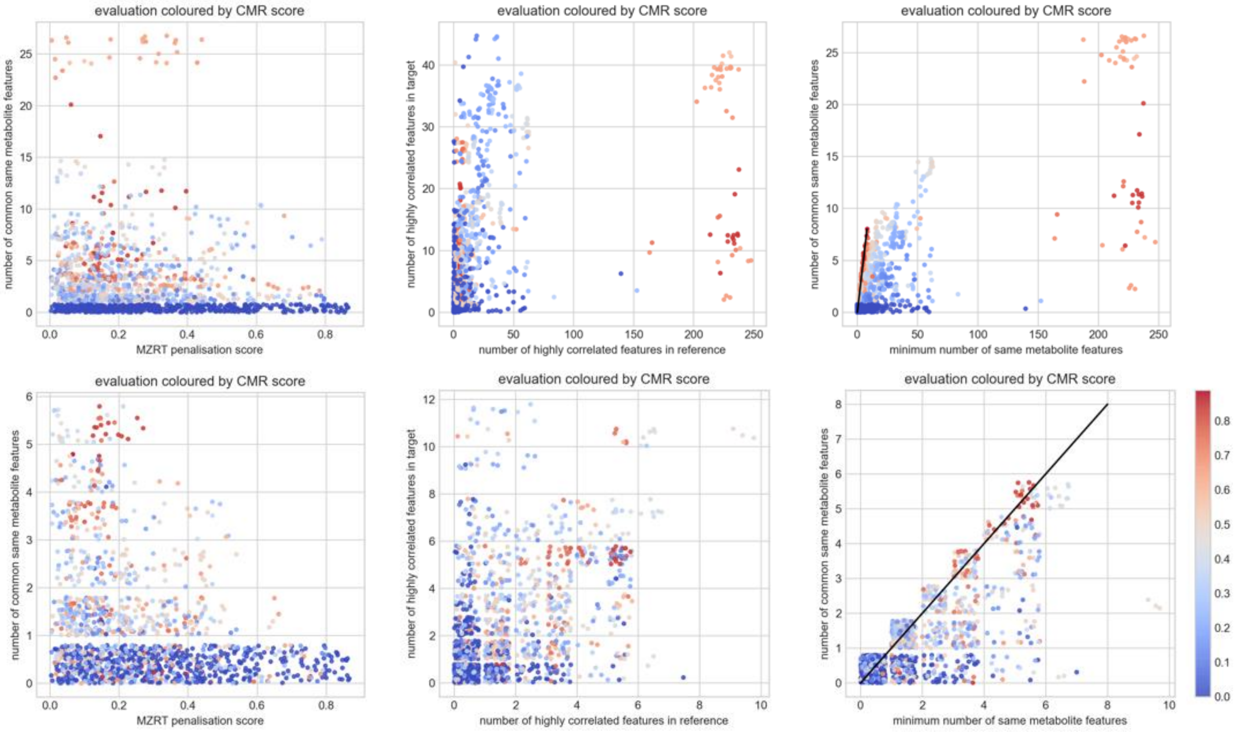


Figure 19. Evaluation of the matching results of MESA 1 and MESA 2.

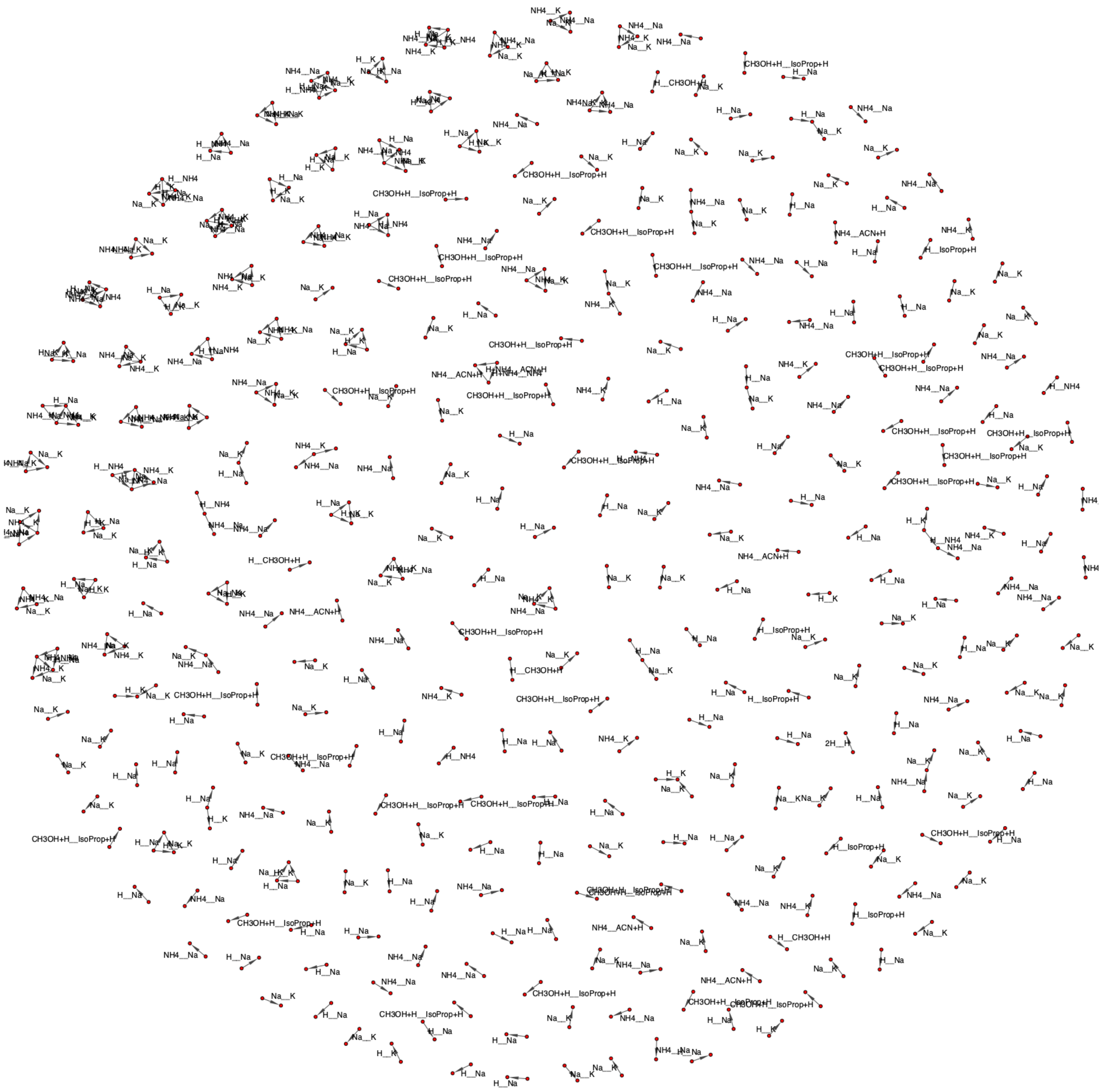


Figure 20. Larger size of figure 8 (a).

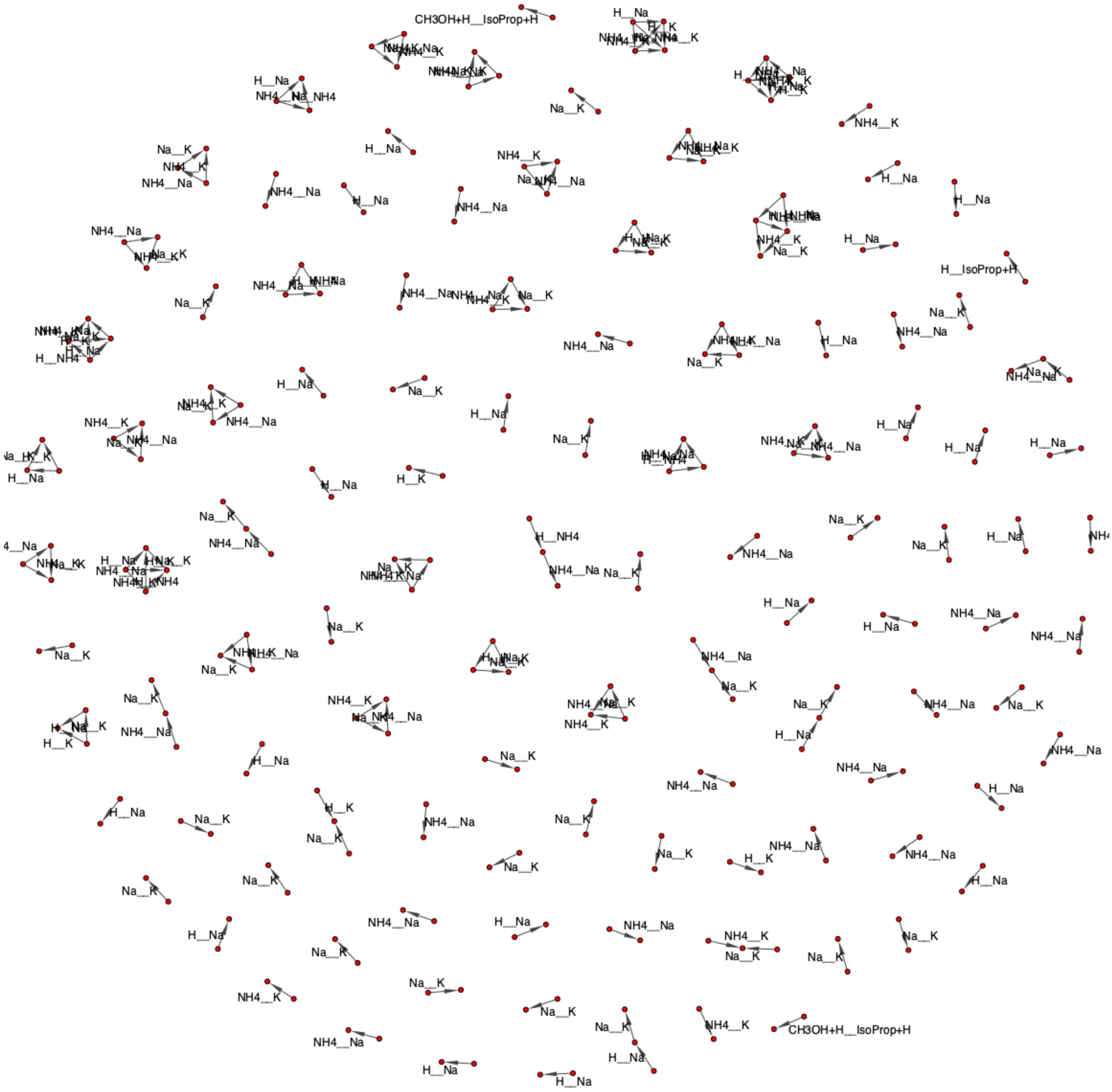


Figure 21. Larger size of figure 8 (b).

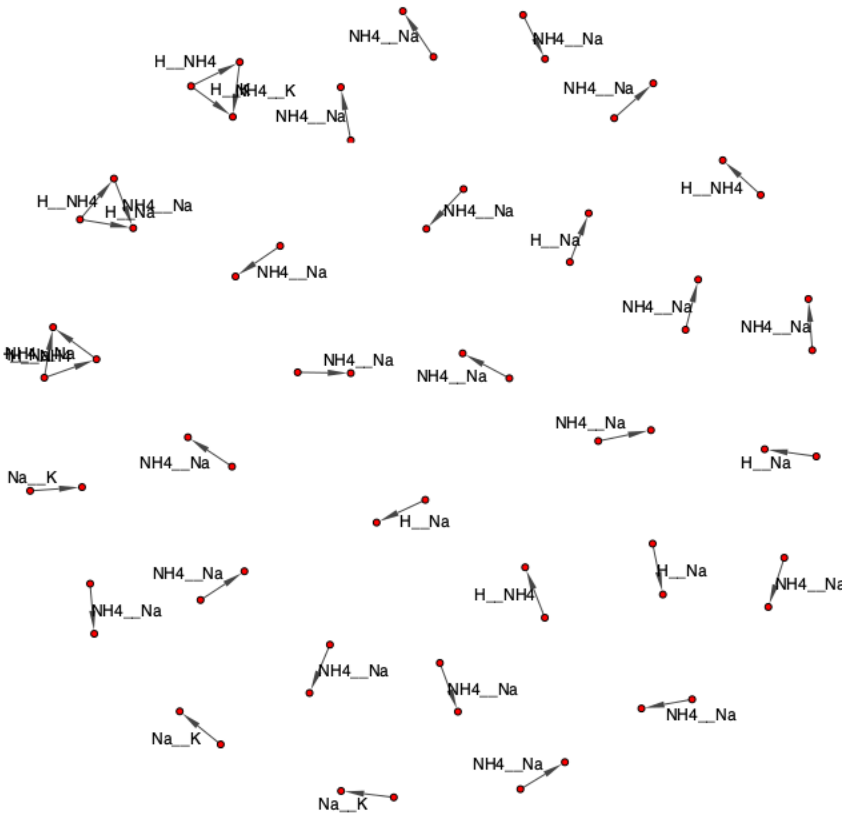


Figure 22. Larger size of figure 8 (c).