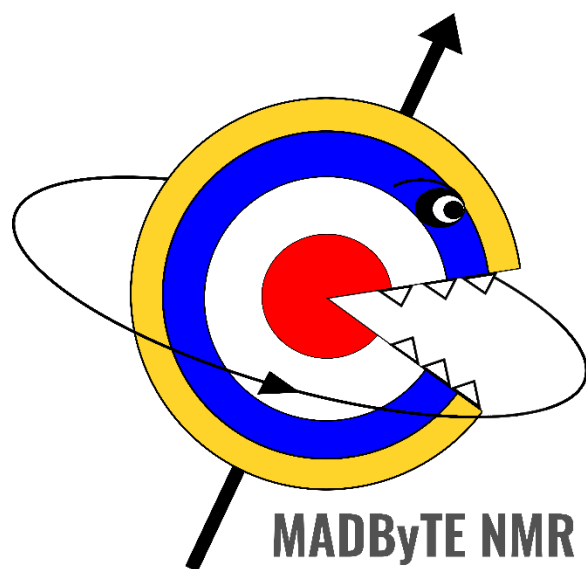


MADByTE Users Guide

Manual Version 05.04.2020



©2020 Joseph M. Egan + MADByTE NMR

Contents

1: Setup	4
1.01: Installation	4
1.02: Prerequisites And Data Processing	4
Bruker and Topspin Processing	4
MestreNova Processing	4
2: MADByTE Comparisons	5
2.01: How MADByTE Works.....	5
2.02: Getting Started.....	6
2.03: Running MADByTE	6
Running MADByTE on a subset of data:	6
Select NMR data directory	7
Select NMR Data Type:	8
Select Project Directory:	8
2.04: Processing Parameters and Filtration Cutoffs	9
Hppm_Error	9
Cppm_Error:.....	9
Consensus_Error:	10
Similarity_Ratio:.....	10
3: Networks.....	10
3.01: Generating Networks.....	10
Network Structure	10
Types of Networks	11
3.02: Viewing And Manipulating Networks	12
3.03: The Network Viewer Plugin	13
3.04: Customizing your network.....	13
3.04: Bioactivity Mapping	14
Bioactivity Data Format	14
Mapping the bioactivity to the network.....	15

4: Dereplication.....	17
4.01: Using The Dereplication Library.....	17
4.02: Establishing a Dereplication Library.....	18
To submit a peak list as a reference for dereplication	19
To Add New Data Into The Dereplication Library Directly.....	19
4.03: Other Dereplication Options – Integration with SMART NMR	20

1: Setup

1.01: Installation

At this time, we will help directly with installation, so this section is largely for future implementation. For now, take solace in the fact that we're here to help.

1.02: Prerequisites And Data Processing

MADByTE uses the peak picked data from Bruker's Topspin or MestreNova to construct the features used in the comparison. This means that the user must perform all the necessary processing – linear prediction, phasing, peak picking – that they wish before running MADByTE.

Once your data processing is complete, simply follow the export steps listed below.

Bruker And Topspin Processing

MADByTE uses peak picked data directly from Bruker without the need to organize it in a special directory – however, it must be exported from Bruker correctly. The recommended export is to run the command 'convertpeaklist txt' from the input bar in topspin. This command exports all the information contained in the peaks tab of a given experiment out as a .txt file within the pdata folder for that experiment without needing to create new directories for your data.

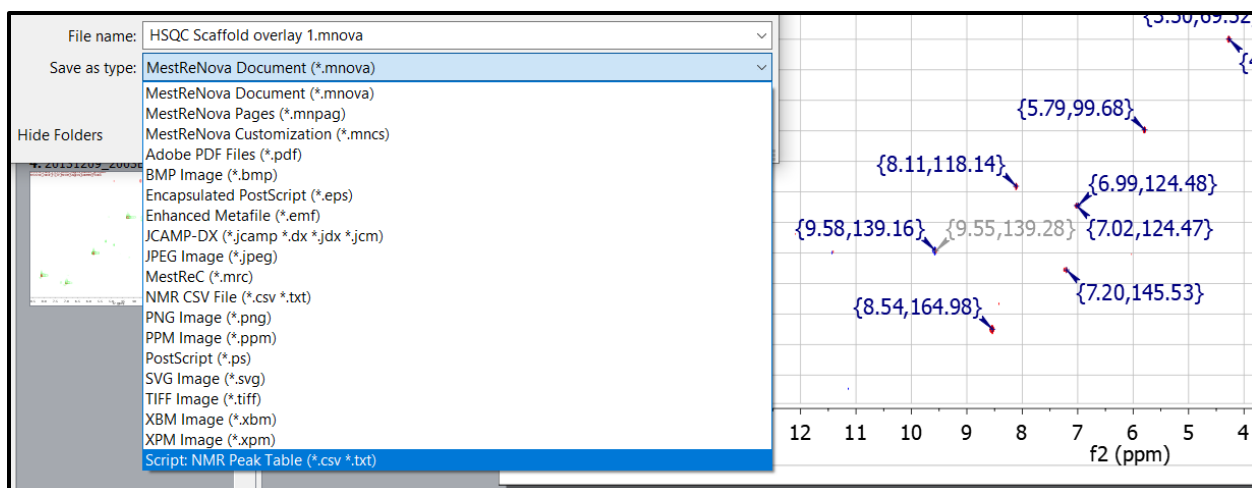
MestreNova Processing

MestreNova exporting is slightly more complicated than that of the Topspin method, but gives the users access to a platform independent method of processing their data.

Once the data has been peak picked, simply go to file, export as, and select Script: NMR Peak Table and format the name as:

[Sample_Name_Here]_HSQC.csv or [Sample_Name_Here]_TOCSY.csv

Place both of these files in a directory named for the sample.

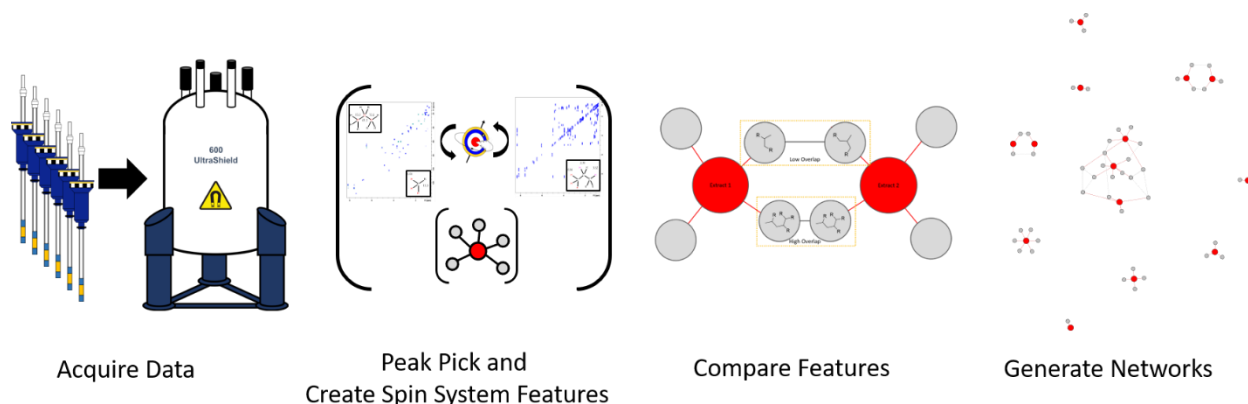


A suggested data directory structure would be:

- MADByTE Input Files
 - [Experimental_Set_Name]
 - [Sample_Name]
 - [Sample_Name_HSQC.csv]
 - [Sample_Name_TOCYSY.csv]

2: MADByTE Comparisons

2.01: How MADByTE Works



MADByTE works by taking the peak picked information and attempting to derive spin systems with carbon correlations. In essence, it is a way to use the scalar coupling information provided by TOCSY/COSY and linking it to the HSQC heteroatom information. This way, if proton resonances fluctuate due to changing substituents, you still match the core scaffold pieces. Additionally, solvent conditions, matrix contribution, etc... all of these have an effect on chemical shift which would cause simple matching to fail. However, by

using an orthogonal viewpoint, these shifts can be perturbed and still provide extremely important information.

2.02: Getting Started

Once the NMR data has been peak picked and exported using the proper output formats, MADByTE is ready to be run. There are a few considerations to take into account that will affect the outcome of the processing, listed at the end of this section.

2.03: Running MADByTE

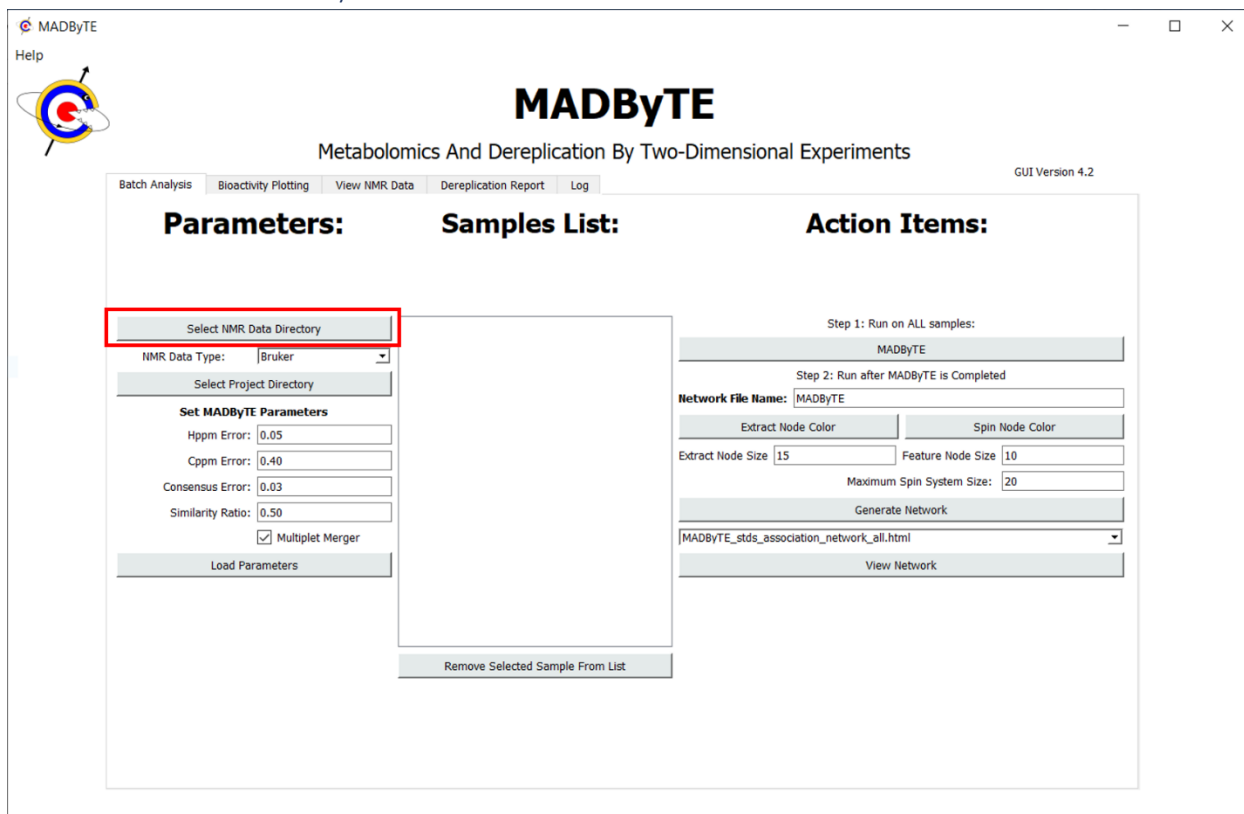
When running MADByTE, you must have already constructed the peak picked lists as outlined above. If you have not done that, do not read further.

Running MADByTE On A Subset Of Data:

To run MADByTE, simply follow the on-screen prompts in the GUI. Each MADByTE processing is treated as completely independent, so if you want to simply add more samples into an older experiment file, or re-process data with new cutoffs, simply re-select the NMR data directory.

As a suggestion, create 'smart' groupings to process. For instance, you may have data on 150+ extracts, but if they're all from different organisms and different chromatographic separations, they may not share any notable metabolites.

Select NMR Data Directory

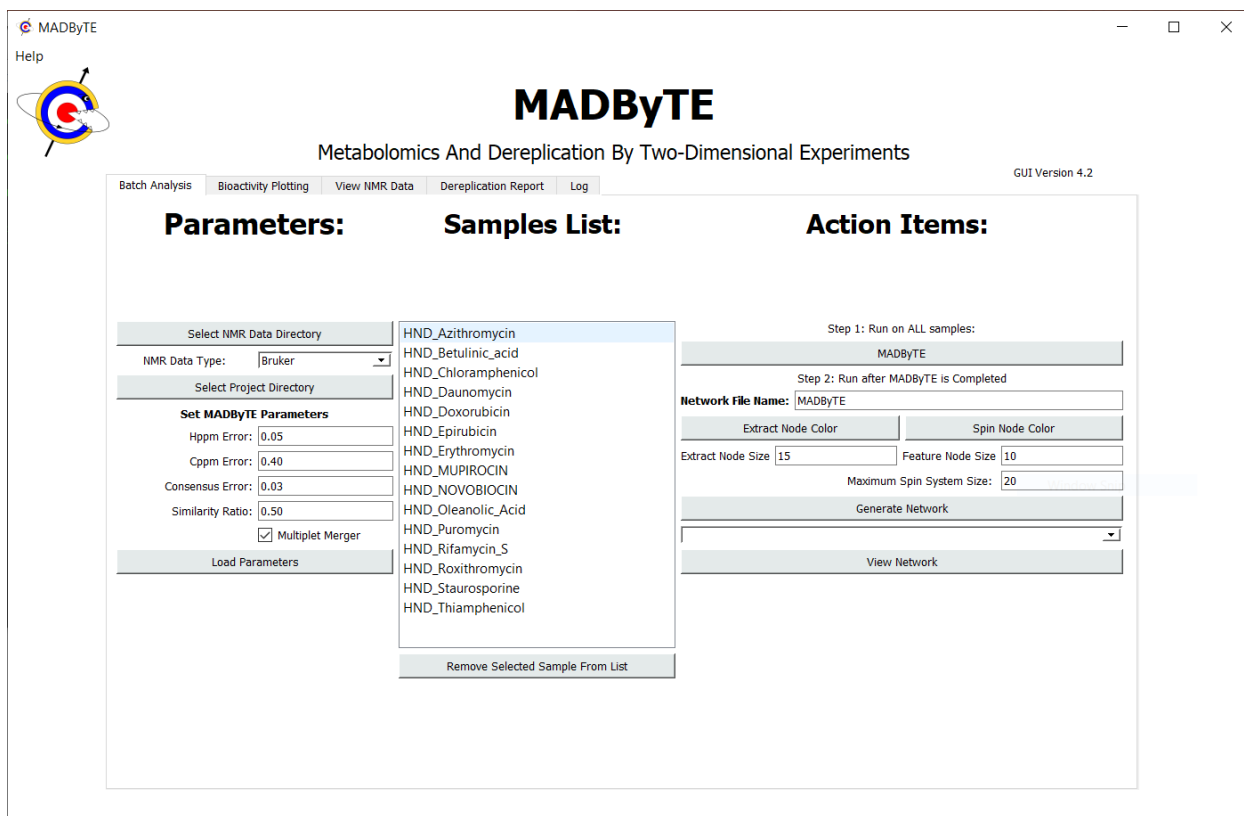


The screenshot shows the MADByTE GUI with the following components:

- Header:** MADByTE logo, title "Metabolomics And Dereplication By Two-Dimensional Experiments", and version "GUI Version 4.2".
- Navigation:** Tabs for "Batch Analysis", "Bioactivity Plotting", "View NMR Data", "Dereplication Report", and "Log".
- Parameters:**
 - Select NMR Data Directory:** A button highlighted with a red box.
 - NMR Data Type:** A dropdown menu set to "Bruker".
 - Select Project Directory:** A button.
 - Set MADByTE Parameters:**
 - Hppm Error: 0.05
 - Cppm Error: 0.40
 - Consensus Error: 0.03
 - Similarity Ratio: 0.50
 - ☒ Multiplet Merger
 - Load Parameters:** A button.
- Samples List:** A large empty box for displaying sample data.
- Action Items:**
 - Step 1: Run on ALL samples:** A button labeled "MADByTE".
 - Step 2: Run after MADByTE is Completed:**
 - Network File Name:** A text field containing "MADByTE".
 - Extract Node Color:** A button.
 - Spin Node Color:** A button.
 - Extract Node Size:** A text field containing "15".
 - Feature Node Size:** A text field containing "10".
 - Maximum Spin System Size:** A text field containing "20".
 - Generate Network:** A button.
 - View Network:** A button.

This is where your RAW data is kept. In the case of Bruker, this will be your NMR data folder that contains all the directories topspin uses. In the case of Mestrenova, this will be the directory you formatted to contain your peak picked lists (See above section – [Mestrenova Data](#)).

When you select the directory, you will see the sample list populated:



If you have a sample in this list that you do not wish to run, simply highlight it and select ‘remove sample’. If you accidentally remove a sample you wanted to keep, simply re-select the NMR data directory.

Select NMR Data Type:

MADByTE does it’s best to be automated, but you must select the data type to process. In the drop down list, simply select your data type. Currently, we support Mestrenova peak picked lists and Topspin peak picking outputs. Do you use something else and want to see it supported? Get in touch with me, and let’s get it working!

Select Project Directory:

This is the most important section to remember to keep track of. When MADByTE is finished processing, all the data derived, the graphs you will eventually generate, and the correlation matrix will be stored in this folder. So, give it a descriptive name and remember you can always delete folders later.

As a suggestion, if you were running a sample of 20 extract prefractions that were hits in an assay against MSRA, create a folder that describes that like “MADByTE_Analysis_Of_MRSA_Hits_date”

If you are using MADByTE to layer networks – see [bioactivity layering](#) - or if you are hoping to view the NMR data in the MADByTE plotting tool, you will need to select this data directory to manipulate the data.

2.04: Processing Parameters and Filtration Cutoffs

There are a few important cutoffs defined in the GUI which will have a series of downstream effects in the processed data.

Hppm_Error

The Hppm_Error is defined as how far two points can be from one another in the 1H dimension and be considered the same point. This is in reference to **both** the construction of the spin systems, as well as the matching of the spin systems to each other. Therefore, the tighter the restrictions, the better the data will agree in the final networks – however, too tight of restrictions will cause there to be no matches in the final networks.

To establish a basis for comparison, the maximum value of a scalar coupling was taken into consideration and assumed to be observable in an imperfect HSQC. Such that, you still see multiplet structure in an HSQC correlation. If this were the case, then a value of around 16Hz would be the maximum scalar coupling according to the standard Karplus plot. When converting the Hz to ppm for a 600MHz magnet, we find that a value can shift as much as 0.026ppm. Taking into account that this can happen in either direction (or in both directions in the case of a triplet), we double this threshold to be 0.05ppm. This is an estimate, and works well for most applications. To establish your cut-offs, simply use the suggested equation below:

$$2 \times \left(\frac{\text{Maximum Hz Splitting Allowed (Hz)}}{\text{Carrier Frequency (MHz)}} \right) = \text{Hppm_Error}$$

And using a 600MHz magnet as a basis for this example:

$$2 \times \left(\frac{16\text{Hz}}{600\text{MHz}} \right) = 0.05\text{ppm} = \text{Hppm_Error}$$

Cppm_Error:

Using the same logic as the above Hppm_Error, the Cppm_Error is how far off a carbon resonance can be off to be considered the same. This restriction is more important when looking for analogs, as small changes in the scaffold can have a bigger effect on the Cppm values than one would expect. However, as a suggestion, the same equation above can be used to determine the Cppm_Error cutoff, but remembering that the gyromagnetic ratio for ¹³C is ¼ that of ¹H, remember to adjust the carrier frequency.

$$2 \times \left(\frac{\text{Maximum Hz Splitting Allowed (Hz)}}{\text{Carrier Frequency (MHz)}} \right) = \text{Cppm_Error}$$

Using an example frequency of 125Hz, we find that our values should be

$$2 \times \left(\frac{125\text{Hz}}{150} \right) = 1.6\text{ppm} = \text{Cppm_Error}$$

However, practically, this is a very wide window. So, for many of the applications that MADByTE was piloted on, we use 0.4ppm as the error window.

Consensus_Error:

The Consensus_Error parameter is how close a resonance must be in the HSQC and TOCSY to be considered the same proton resonance. In a perfect world, this would be an exact match. However, running TOCSY pulse sequences tends to heat the sample a little, which can cause minor perturbations in the chemical shifts. As a generalizable parameter, we find that working with between 0.03ppm and 0.05 is a good starting point.

Similarity_Ratio:

The similarity ratio is used after the calculation of the correlation matrix. This means that when all of the spin systems are defined, it compares the membership of each list to each other list based on the Hppm_Error and the Cppm_Error. These ratios are stored in the correlation matrix. When used to network things together. The way the networking step works is that it fetches all correlations that the similarity score (value in the correlation matrix) is higher than the Similarity_Ratio cutoff.

So, considering a simple case:

Spin_System_1 = [(1.00,18.0),(2.15,24.6),(8.4,127.5)]

Spin_System_2 = [(1.04,18.2),(2.18,24.8),(8.6,129.0)]

Using a Cppm_Error of 0.4ppm and an Hppm_Error of 0.05, a pairwise comparison would yield:

	(1.00,18.0)	(2.15,24.6)	(8.6,129.0)
(1.04,18.2)	MATCH	NO MATCH	NO MATCH
(2.18,24.8)	NO MATCH	MATCH	NO MATCH
(8.4,127.5)	NO MATCH	NO MATCH	NO MATCH

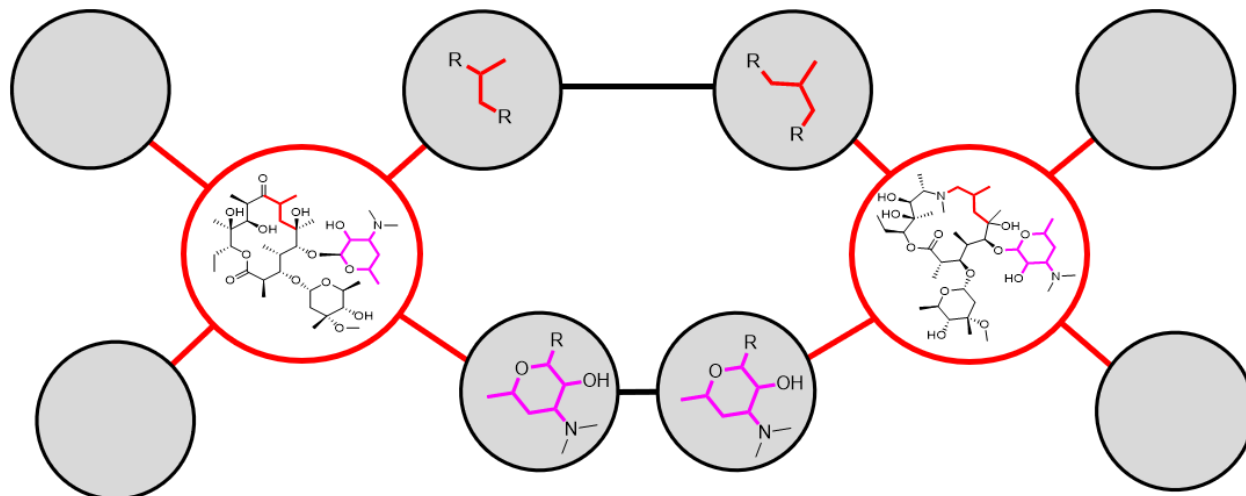
This means that 2/3 of the resonances show a match, yielding a similarity ratio between the spin systems of 0.66 (or 66%). If the similarity ratio was set to 0.5 (50%), they would link together in the network.

3: Networks

3.01: Generating Networks

Network Structure

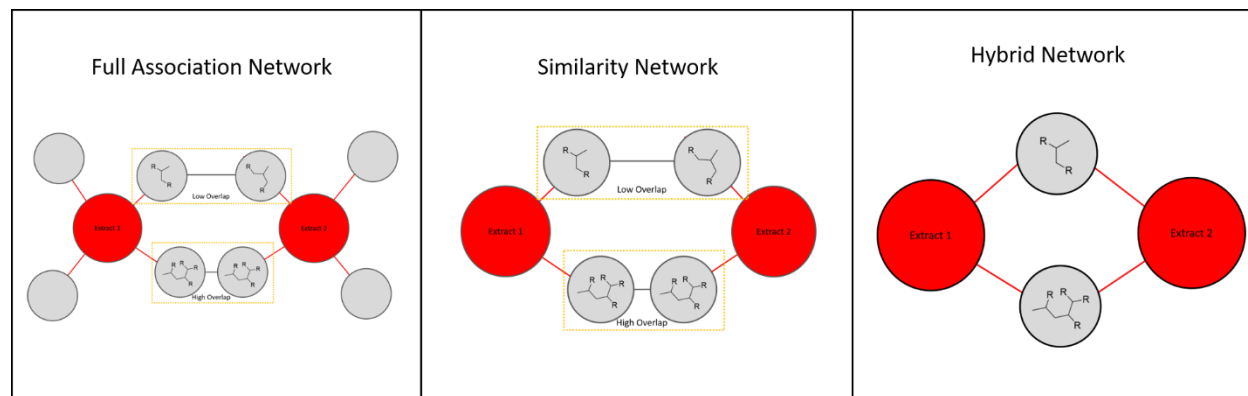
MADByTE Networks are collections of sample nodes and spin system feature nodes, tied together by their comparison scores. MADByTE looks for common proton and carbon resonances in the constructed spin systems and gauges the amount of overlap between two spin system features. If the overlap is greater than the similarity cutoff, then the spin systems are connected to each other.



In most networks, the strength of the connection is equal to the amount of overlap between the two spin systems. This means that if you have the same chemical motif conserved across several molecules (such as the desoamine moiety in azithromycin and erythromycin), the overlap will be very high, and the nodes should have more weight between them.

Types Of Networks

MADByTE outputs three different types of networks, each with their own purpose.



Full Association Network

The Full Association network is a comparison network that renders all nodes for every spin system detected in every sample. After this, it draws edges between nodes displaying spectral overlap that is

greater than the similarity ratio cutoff, and the weight between them is proportional to the similarity ratio between them.

Similarity Network

The similarity network is a reduced complexity mapping of the shared nodes **only**. If an extract or compound does not have any spin systems shared with the rest of the data set, neither its spin systems nor its extract node will be rendered. This enables a rapid viewpoint of the most conserved chemical motifs within the sample subset.

Hybrid Network

The Hybrid Network is a rendering of the Similarity Network that combines the connected nodes into their representative shared chemistry. The result will be a list of all points seen in all the similar nodes.

3.02: Viewing And Manipulating Networks

Networks can be viewed directly in the MADByTE system by navigating to the main “Batch Analysis” screen and selecting the network from the drop down list and selecting ‘view network’.

- If the dropdown list is empty, simply select the project directory you wish to query and the networks contained in that folder will populate the drop down list.

The screenshot displays the MADByTE GUI interface. At the top, the title bar reads "MADByTE" and the subtitle is "Metabolomics And Dereplication By Two-Dimensional Experiments". The main window is divided into three sections: "Parameters:", "Samples List:", and "Action Items:". The "Parameters:" section includes fields for "Select NMR Data Directory", "NMR Data Type" (set to "Bruker"), "Select Project Directory", and "Set MADByTE Parameters" with input fields for "Hppm Error" (0.05), "Cpm Error" (0.40), "Consensus Error" (0.03), "Similarity Ratio" (0.50), and a checked "Multiplet Merger" checkbox. The "Samples List:" section is currently empty. The "Action Items:" section contains a "Step 1: Run on ALL samples:" button labeled "MADByTE", followed by a "Step 2: Run after MADByTE is Completed" section. This section includes a "Network File Name:" dropdown menu (currently showing "MADByTE"), "Extract Node Color" and "Spin Node Color" buttons, "Extract Node Size" (15) and "Feature Node Size" (10) input fields, and a "Maximum Spin System Size" (20) input field. Below these is a "Generate Network" button. A red box highlights the "Network File Name:" dropdown menu, which shows a list of network files, including "MADByTE_stds_association_network_all.html". Below the dropdown is a "View Network" button.

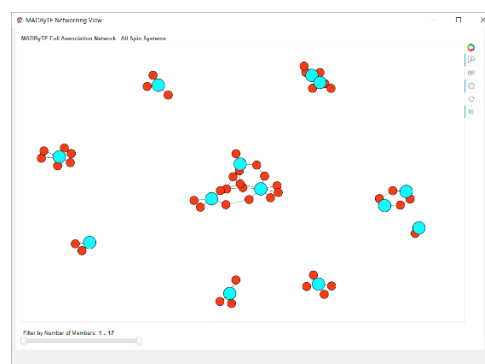
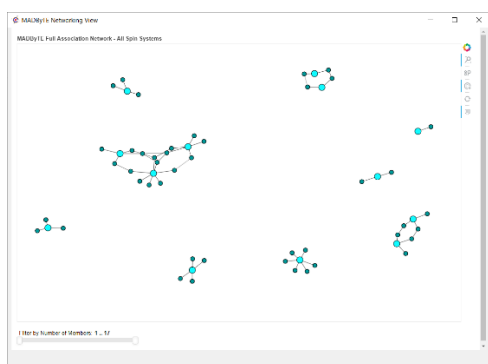
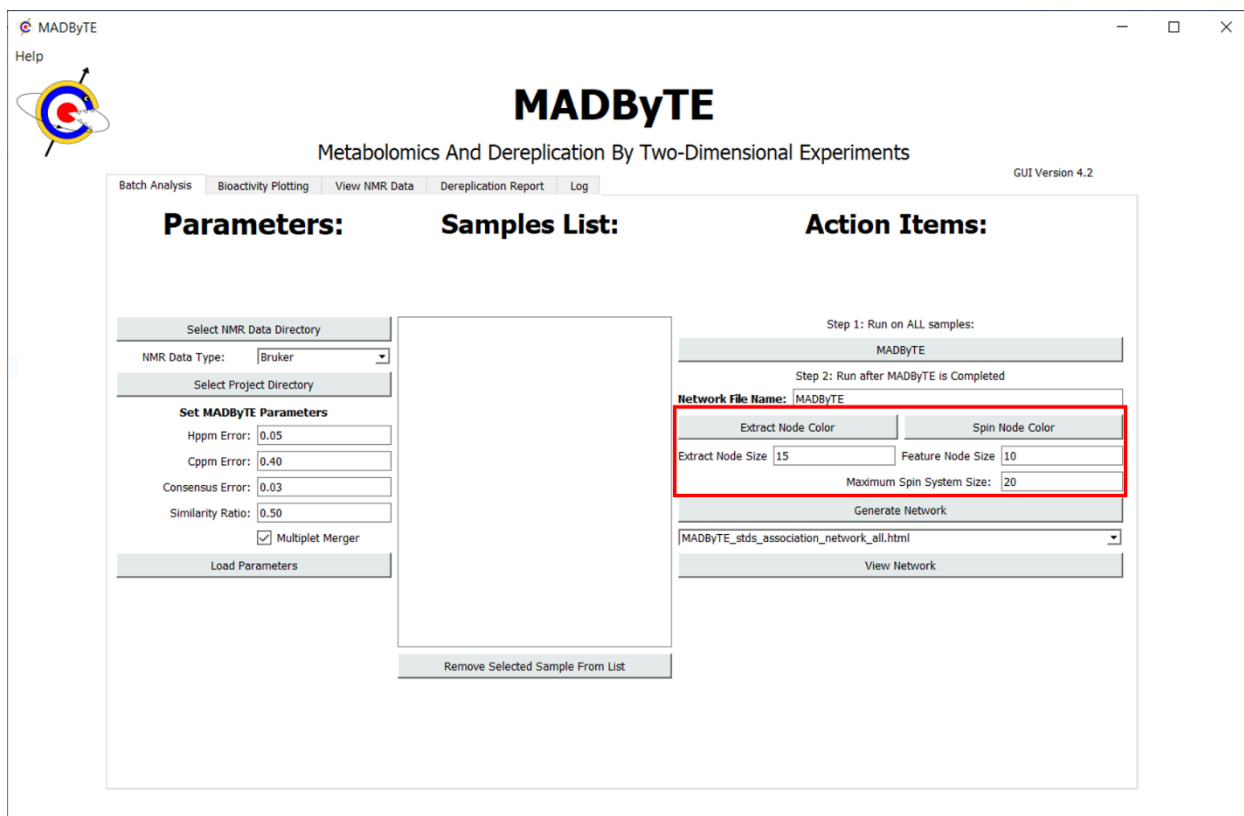
3.03: The Network Viewer Plugin

MADByTE's native network viewer renders the HTML files generated by MADByTE in a pop up window for easy navigation (Shown below – left). To filter the network by spin system size, simply drag the bar at the bottom of the screen to represent your requirements and the GUI updates automatically(Shown below – right).



3.04: Customizing Your network

Networks generated by MADByTE can be adjusted based on your personal preferences. Color and size of the nodes can be adjusted simply by changing the values in the GUI. The standard layout is shown on the bottom left, and an alternate size and color scheme is shown on bottom right.



3.04: Bioactivity Mapping

MADByTE contains a module for mapping bioactivity evaluations directly onto the sample ID nodes as a way to prioritize samples that show high bioactivity and shared structural motifs. To map the data correctly, there are a few things that must be considered.

Bioactivity Data Format

MADByTE allows for the layering of a single assay result on top of the sample ID nodes, but to do this it requires that the sample names match the NMR sample names. As an example, if sample "JE_Erythromycin" is the name of the NMR sample, it must also be named "JE_Erythromycin" in the bioactivity file. This means that dose responses **are not** factored in at this time.

The data must be saved as a CSV in the following format:

	A	B
1	Sample	Bioactivity_Score
2	HND_Azithromycin	1
3	HND_Betulinic_acid	0
4	HND_Chloramphenicol	0
5	HND_Daunomycin	0
6	HND_Doxorubicin	0
7	HND_Epirubicin	0
8	HND_Erythromycin	1
9	HND_Mupirocin	0
10	HND_Novobiocin	0
11	HND_Oleanolic_Acid	0
12	HND_Puromycin	0.5
13	HND_Riffamycin_S	0
14	HND_Staurosporine	0
15	HND_Thiamphenicol	0.5
16	HND_Roxithromycin	1

Notice that the scale here is from 0-1, meaning that you must normalize your data to fit a linear scale. It **does not need to be** from 0-1, but the values will be important in the next step.

Mapping The Bioactivity To The Network

To map your bioactivity profile to your NMR data, simply navigate to the bioactivity plotting tab of the GUI and walk through the steps outlined.

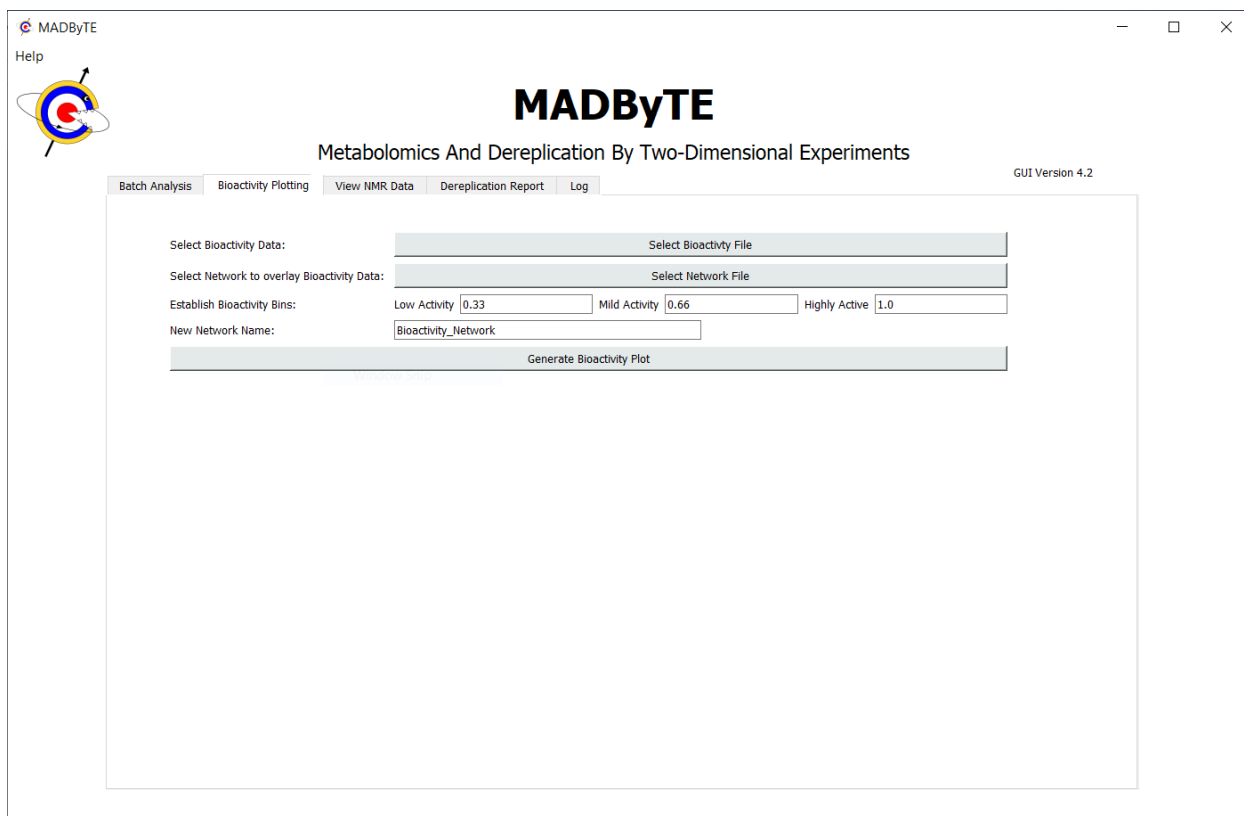


Figure 1: The Bioactivity Plotting Window

Select The Bioactivity File

The bioactivity profile must be in CSV format, as outlined above.

Select Network File

The Network File should be the graphml of the network that you wish to overlay. The bioactivity overlay can work with any of the networks generated by MADByTE.

Establish Bioactivity Bins

By default, MADByTE assumes a scale of bioactivity that has been normalized from 0-1. However, if your bioactivity is measured on a different scale, simply redefine the bins according to the bioactivity you wish to highlight. You can generate many networks that overlay different types of bioactivity data by simply renaming the file and regenerating the network.

Name The Network

The resulting network will be saved as an HTML file that contains the new bioactivity highlights.

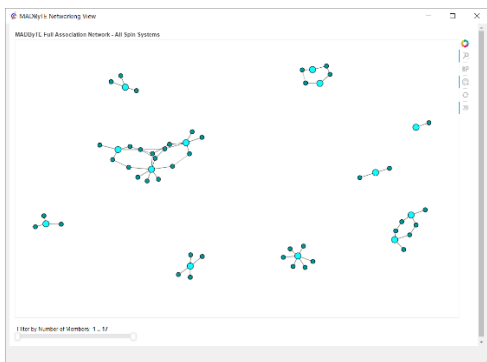


Figure 2: A Typical Network From MADByTE

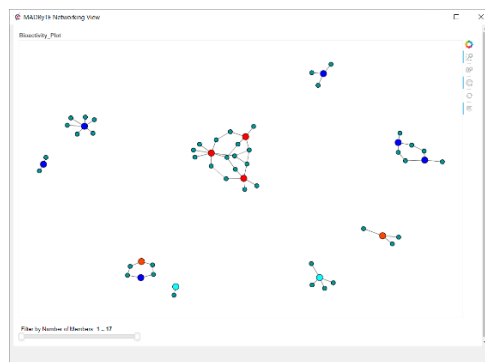


Figure 3: Bioactivity Color Coding Using The Bioactivity Integration Option

The “low” bioactivity compounds are color coded as blue, the ‘mild’ bioactivity compounds are color coded as orange, and the ‘highly’ bioactive compounds are color coded as red. In this instance, there were two compounds that the bioactivity data is missing for. For these compounds, the default color coding is retained.

4: Dereplication

Dereplication, or predicting known molecular entities present in a mixture using reference data is possible with MADByTE using the Dereplication module. Using a database established by the user, molecular entities familiar to the user can be stored as reference files and compared against any data that has been processed by MADByTE, eliminating the need for isolation under certain circumstances.

4.01: Using The Dereplication Library

To use dereplication, navigate to the Dereplication Report tab on the MADByTE GUI. There, you will see a sample drop down list populated by the samples you have processed through MADByTE previously.

- Select the sample you wish to query against the dereplication database
- Fill in the error margins you wish to use.
- Click Dereplciate and the results will be populated in the window to the right

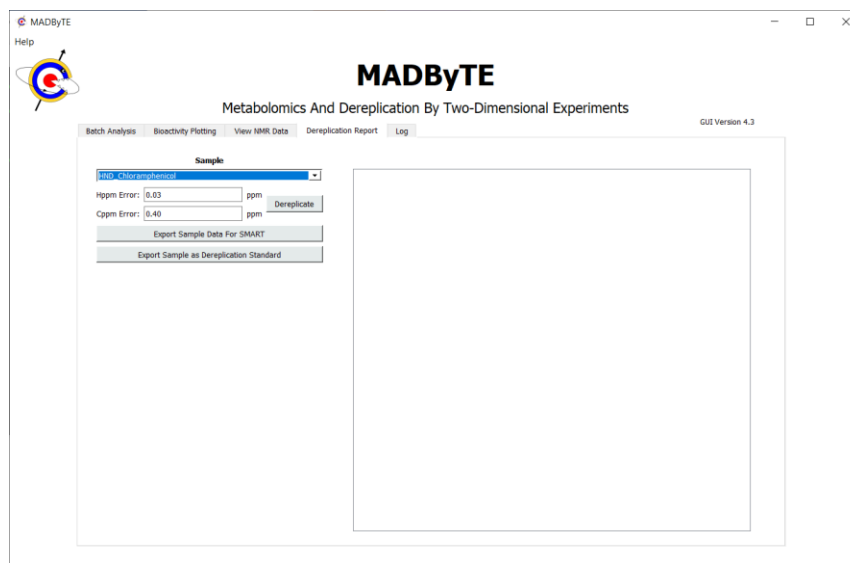


Figure 4: Dereplication Report Tab

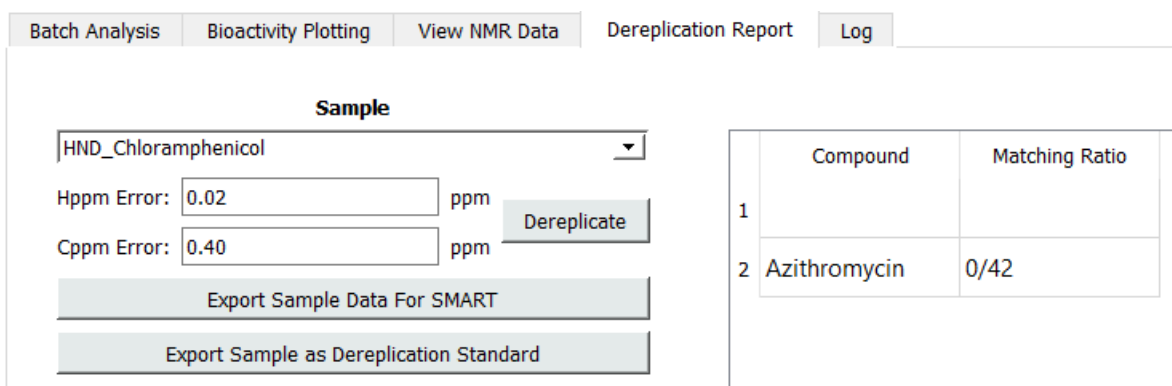


Figure 5: Dereplication Output and Options

4.02: Establishing A Dereplication Library

A major hurdle to effective dereplication is the lack of a curated and open centralized repository for data to compare against. Further, a centralized server may be a security risk for some applications or users who wish to keep their data private and on a local machine. To facilitate dereplication under these circumstances, MADByTE offers the users a way to establish a local dereplication library which they themselves control and allow for curation.

Users can submit peak lists created by themselves, or from simulated data by storing them in the directory called Dereplication_Database within the MADByTE program folder. Once a sample has been processed by MADByTE, the system will be able to query the sample for signals that match reference compounds within user definable margins.

To Submit A Peak List As A Reference For Dereplication:

1. Peak pick the HSQC data for your pure compounds.
2. Save the peak picked data as a csv file with the format:
 - 2.1. H_ppm, C_ppm, Identity
 - 2.1.1. Report H_ppm to 2 decimal places
 - 2.1.2. Report C_ppm to 2 decimal places
 - 2.1.3. List the ID of the compound in the third column throughout the list
 - 2.2. Convert the csv to json
 - 2.3. Save the file as: DDF_[name of compound here].json
3. Place the file in the folder called Dereplication_Database, and you're done!
4. If you have legacy data from old publications, literature, or an in-house archive, simply format it as a json table and use the same naming convention in step 2.3.

Several sample compounds are included as a default database to get started.

To Add New Data Into The Dereplication Library Directly:

MADByTE has an integrated method to directly submit sample data into the dereplication library. Simply process the data using the MADByTE analysis, and then navigate to the Dereplication Report tab in the MADByTE GUI.

Select the sample you wish to submit into the dereplication library and click 'Export Sample as Dereplication Standard'. This will output a list of the peak picked data used for MADByTE analysis directly into the Dereplication_Database folder and add it to the list of compounds to be searched against during the next dereplication event.

Sample	
HND_Chloramphenicol	
Hppm Error: 0.02	ppm
Cppm Error: 0.40	ppm
<button>Dereplicate</button>	
<button>Export Sample Data For SMART</button>	
<button>Export Sample as Dereplication Standard</button>	

	Compound	Matching Ratio
1		
2	Azithromycin	0/42

Figure 6: How to export the sample as a dereplication database file

Batch Analysis
Bioactivity Plotting
View NMR Data
Dereplication Report
Log

Sample

HND_Chloramphenicol

Hppm Error: 0.02 ppm
Cpmm Error: 0.40 ppm
Dereplicate

Export Sample Data For SMART

Export Sample as Dereplication Standard

	Compound	Matching Ratio
1		
2	Azithromycin	0/42
3	HND_Chloramphenicol	8/8

Figure 7: Running the sample again after submitting a dereplication database file yields new results in the window

4.03: Other Dereplication Options – Integration With SMART NMR

Although there is no current standard for submission of NMR data into a centralized server, some groups have constructed fantastic resources that streamline dereplication protocols. One of these, known as SMART (Small Molecule Accurate Recognition Technology) is a webserver constructed and maintained by a team at USCD. SMART allows for users to drag and drop their NMR data and use an artificial intelligence based tool to evaluate the spectra and return like molecules. To facilitate this as a dereplication option, we have integrated an export option into MADByTE. Simply click the “Export Sample For SMART” button and a SMART compatible peak list will be put in the data output from MADByTE allowing you to search SMART for molecules of interest.