Chapter 8

PGFinder, an Open-Source Software for Peptidoglycomics: The Structural Analysis of Bacterial Peptidoglycan by LC-MS

Brooks J. Rady and Stéphane Mesnage

Abstract

Peptidoglycan is a major and essential component of the bacterial cell envelope that confers cell shape and provides protection against internal osmotic pressure. This complex macromolecule is made of glycan strands cross-linked by short peptides, and its structure is continually modified throughout growth via a process referred to as "remodeling." Peptidoglycan remodeling allows cells to grow, adapt to their environment, and release fragments that can act as signaling molecules during host-pathogen interactions. Preparing peptidoglycan samples for structural analysis first requires purification of the peptidoglycan sacculus, followed by its enzymatic digestion into disaccharide peptides (muropeptides). These muropeptides can then be characterized by liquid chromatography coupled mass spectrometry (LC-MS) and used to infer the structure of intact peptidoglycan sacculi. Due to the presence of unusual crosslinks, noncanonical amino acids, and amino sugars, the analysis of peptidoglycan LC-MS datasets cannot be handled by traditional proteomics software. In this chapter, we describe a protocol to perform the analysis of peptidoglycan LC-MS datasets using the open-source software PGFinder. We provide a step-by-step strategy to deconvolute data from various mass spectrometry instruments, generate muropeptide databases, perform a PGFinder search, and process the data output.

Key words Mass spectrometry, Peptidoglycan, Software, LC-MS, Data deconvolution

1 Introduction

Peptidoglycan is an essential component of the bacterial cell envelope. This single, giant molecule surrounds the cytoplasmic membrane and confers cell shape and resistance to internal turgor pressure. The polymerization of peptidoglycan is the target of the most widely used family of antibiotics, the beta-lactams, and its biosynthetic pathway has been extensively studied since the late 50s. While we know a great deal about the enzymes involved in the assembly of peptidoglycan, the structure of this complex macromolecule remains poorly understood. Peptidoglycan's size and

Frédérique Lisacek (ed.), *Protein Bioinformatics*, Methods in Molecular Biology, vol. 2836,
https://doi.org/10.1007/978-1-0716-4007-4_8,

heterogeneous composition make studying its structure challenging. Peptidoglycan is built from relatively simple disaccharide-peptides; however, once these building blocks have been produced in the cytoplasm, they can be incorporated into the existing peptidoglycan network via distinct cross-linking reactions, and both glycan chains and peptide stems are subject to limited cleavage and modification. These processes, collectively referred to as "remodeling," account for the structural diversity of peptidoglycan (Fig. 1).

The method for analyzing peptidoglycan structure has remained largely unchanged since it was first described in the late 80s [1]. Peptidoglycan purification is straightforward and achieved by boiling bacterial cells in 5% SDS before repeatedly washing in Milli-Q water. Contaminants such as nucleic acids, covalently bound proteins, and polymers can be removed via various enzyme and acid treatments whilst leaving the peptidoglycan intact. Pure peptidoglycan sacculi are then digested using a commercially available enzyme (mutanolysin) to generate disaccharide-peptides. These soluble fragments (called muropeptides) are then reduced and analyzed by reversed-phase high-performance liquid chromatography coupled mass spectrometry (henceforth referred to as LC-MS).

The current bottleneck in peptidoglycan structural analysis comes from a dearth of bioinformatic tools for automating the analysis of mass spectrometry data, precluding any consistent and reproducible strategy. Due to the presence of unusual sugars, amino acids, and cross-links between peptide stems, the identification of muropeptides is not easily performed by existing proteomics or glycomics software. The overall search strategy used in peptidoglycan analysis remains poorly described in the literature and in most cases continues to be a manual, time-consuming, and error-prone process. Although some recent studies have described a clear strategy for data analysis [2, 3], these strategies rely on vendor-specific or proprietary software (Waters, Agilent, MATLAB) that are either not suitable to all types of datasets or are difficult to set up and use.

We have recently described the first open-source software (PGFinder) dedicated to peptidoglycan structural analysis [4]. PGFinder is a vendor-neutral tool available through a web-user interface that identifies muropeptides based on the match between their theoretical monoisotopic mass and observed masses in deconvoluted LC-MS datasets. PGFinder can perform either sequential searches with increasing complexity using dynamic databases, limiting the occurrence of identifications based on mass coincidences, or "one off" searches using complex databases created by the user. Although PGFinder does not currently provide the capability to analyze fragmentation data, this represents a first step toward the consistent and reproducible analysis of peptidoglycan structure.

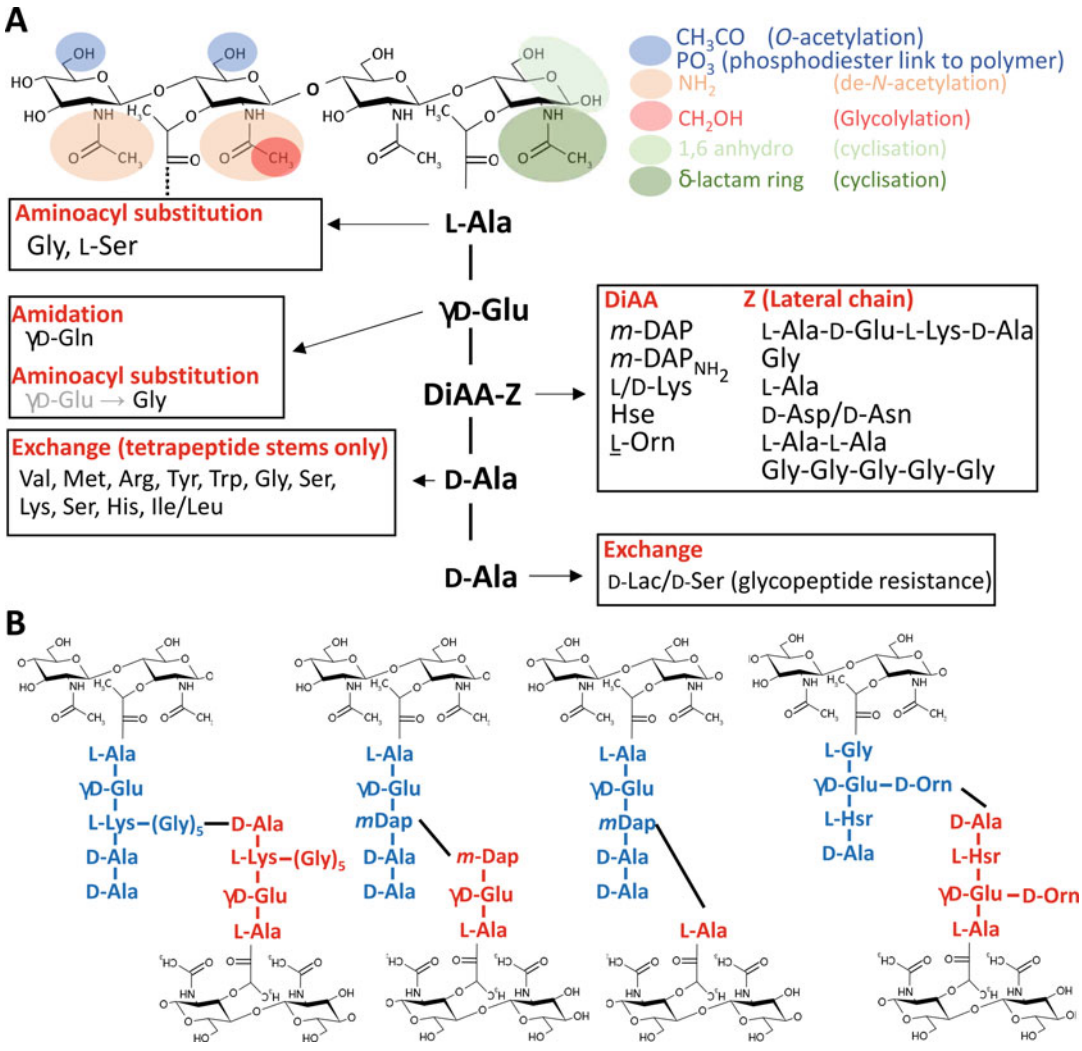


Fig. 1 The diversity of peptidoglycan's composition and structure. (**a**) A representative peptidoglycan building block made of *N*-acetylglucosamine (GlcNAc) and *N*-acetylmuramic acid (MurNAc) forming a disaccharide subunit linked to a pentapeptide stem through the MurNAc's lactyl moiety. The peptide stem contains both L- and D- amino acids and shows a great diversity in composition. Some examples of amino acids found in peptidoglycan are shown for each residue. Modifications of the sugars are also shown. (**b**) A representation of crosslinking diversity; 4-3 bonds (direct or via a peptide crossbridge) are made by D,D-transpeptidases, and 3-3/1-3 bonds are made by L,D-transpeptidases. The enzymes catalyzing 4-2 bonds remain unknown. Acceptor stems are shown in blue, and donor stems in red. DiAA, diamino acid; m-DAP, meso-diaminopimelic acid; D-Lac, D-lactate; Z, lateral chain

In this chapter, we provide a step-by step description of our protocol to analyze peptidoglycan LC-MS data. These steps include data deconvolution, database construction, muropeptide search, and output processing. The strategy presented is applicable to datasets generated by a wide range of mass spectrometry instruments and can be carried out using only open-access software. We

also provide a description of an alternative workflow using the proprietary software platform Byos that is frequently employed by our lab and several others working with mass spectrometry data.

2 Materials

2.1 LC-MS Data

1. LC-MS mass spectrometry data: These can be generated by any instrument, though data generated by ThermoFisher instruments (Orbitrap) and saved as .raw files are the most straightforward to analyze using MaxQuant.
2. (Optional). LC-MS/MS data resulting from data-dependent acquisition: This will allow the user to confirm the matching output. This step is beyond the scope of this review, and MS/MS analysis is not yet supported by PGFinder.

2.2 Software Tools

1. Open-access software tools for converting mass spectrometry data into a standard file format such as mzML [5]. One suggestion is Proteowizard's msConvert [6] found at https://proteowizard.sourceforge.io/download.html
2. Software tools to deconvolute LC-MS datasets. One suggestion is the open-access software MaxQuant [7] found at https://www.maxquant.org/. Another option is the proprietary software Byos® found at https://proteinmetrics.com/byos/
3. Software tool to identify peptidoglycan fragments from deconvoluted LC-MS datasets [4]: PGFinder, found at https://mesnage-org.github.io/pgfinder/
4. Microsoft Excel to visualize search outputs and analyze data, found at https://www.microsoft.com/en-us/microsoft-365/excel. Using an open-source alternative like Libreoffice Calc (https://www.libreoffice.org/discover/calc/) should also be possible, though the figures in this chapter will focus on Excel's user interface and formula syntax

3 Methods

3.1 Data Deconvolution

Before PGFinder can be used to identify muropeptides, the raw LC-MS data must be "deconvoluted." Data deconvolution involves peak picking, deisotoping, and charge assignment that converts raw data into a table of monoisotopic masses, retention times, ion intensity, and other useful parameters. Many software packages may also refer to this process as feature finding, extraction, or detection. LC-MS datasets can be deconvoluted using both the free-to-use MaxQuant and proprietary Byos® packages. MaxQuant

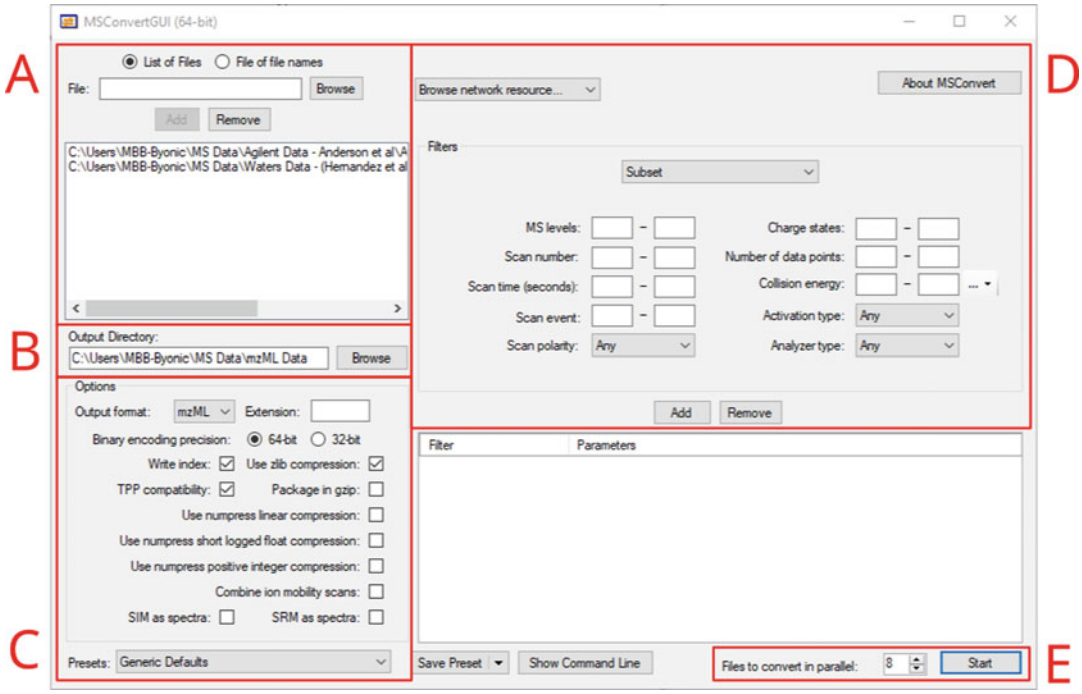


Fig. 2 ProteoWizard's MSConvertGUI

can handle raw data from Thermo Fisher Scientific instruments (.raw files), but datasets generated by other instruments (Bruker Daltonics, AB Sciex, Agilent Technologies, and Waters) may need to be converted to .mzML files before they can be processed for deconvolution. Byos® is vendor-neutral and can handle files from any instrument when deconvoluting data.

3.1.1 Conversion to .mzML Format Using MSConvertGUI

1. After opening MSConvertGUI, ensure the "List of Files" radio button is selected, then select "Browse" to locate the RAW files you'd like to convert. After you have selected some files, click the "Add" button to add them to the conversion queue (Fig. 2a).
2. Next, use the "Browse" functionality to select an output directory for your converted .mzML files (Fig. 2b).
3. Ensure that the selected output format is mzML with 64-bit encoding precision and that the "Write index," "Use zlib compression," and "TPP compatibility" options are selected (Fig. 2c).
4. Optionally, filters can be added to perform operations such as peak-picking, charge-state prediction, and lockmass refining, but none is required for the subsequent steps with MaxQuant, Byos®, or PGFinder (Fig. 2d).

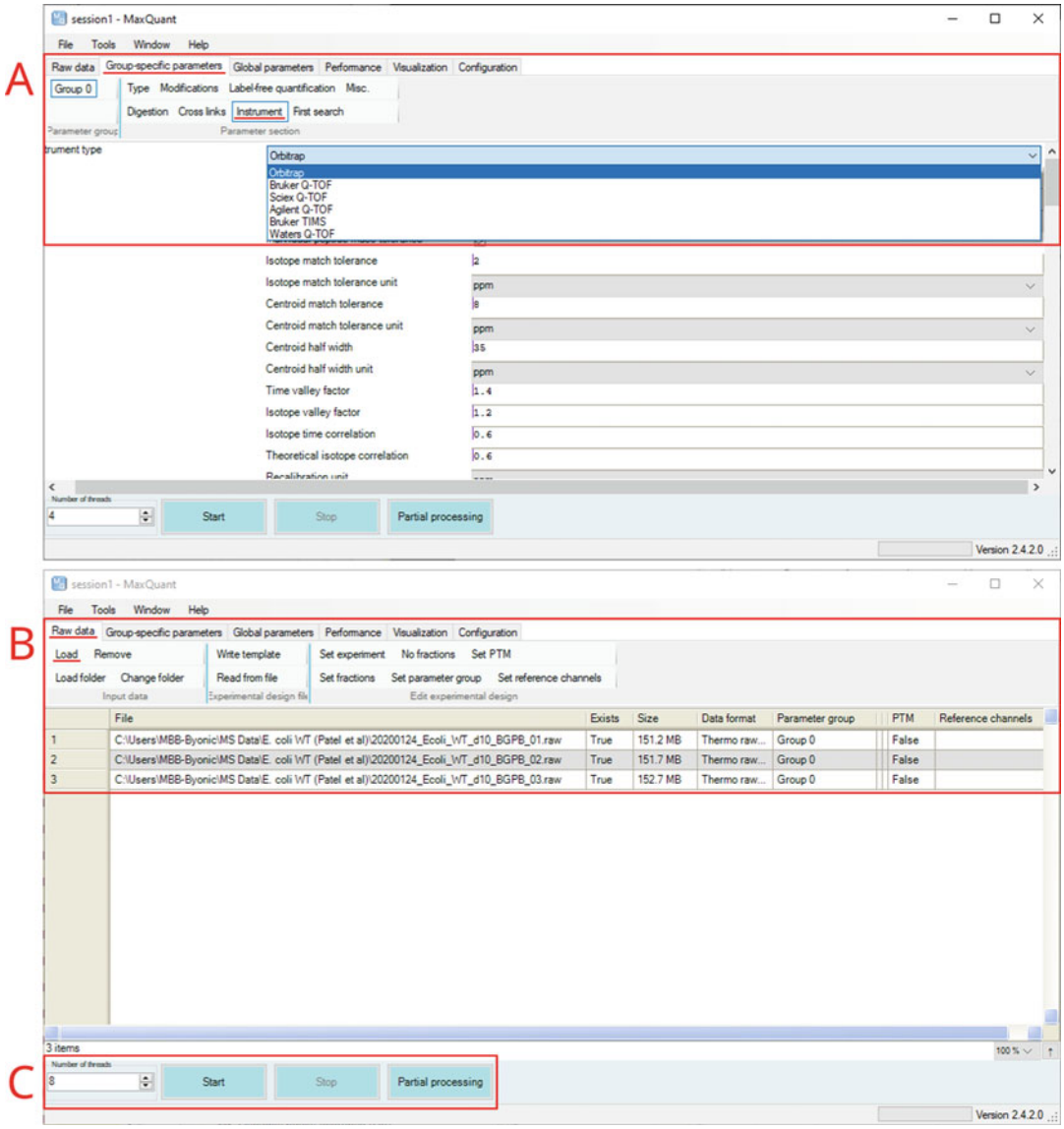


Fig. 3 Configuring MaxQuant for data deconvolution

5. Finally, be sure that you have set the "Files to convert in parallel" option to the number of CPU cores on your computer, and press start (Fig. 2e).

3.1.2 Data Deconvolution Using MaxQuant (.txt File Output)

1. Before processing data using MaxQuant, you need to specify the type of instrument that generated your data. To select your instrument type, first select the "Group-specific parameters" tab, then the "Instrument" sub-tab. From there, you can select your "Instrument type" using the drop-down menu—here, we have selected "Orbitrap" (Fig. 3a).

2. Next, navigate back to the "Raw data" tab and click "Load" to select your data files—either Thermo .raw files or the converted .mzML files (Fig. 3b). Note that selecting several data files at this step will result in them being pooled! It is essential, therefore, to ***only select several files if they are biological or technical replicates of one another!***
3. Finally, set the "Number of threads" option to the number of CPU cores on your computer, and click "Start" (Fig. 3c).
4. After MaxQuant has finished processing your data, navigate to the same directory as your selected input files, and you should find several new folders created by MaxQuant. To find the allPeptides.txt file needed by PGFinder, navigate into the newly created "combined" directory, then into the "txt" directory where you should find the allPeptides.txt file containing tab-separated value (TSV) data describing the monoisotopic masses identified by MaxQuant. You can then copy this file to a safe location for later use by PGFinder.

3.1.3 Data Deconvolution Using Byos® (.ftrs File Output)

1. After opening Byos®, you will be presented with a number of standard workflows, but you can build a custom workflow to generate .ftrs files. Click on the "PTM" workflow icon to select it as a starting point (Fig. 4a).
2. Next, navigate to the "Processing nodes" tab, then click on "Modifications" to reveal the modification settings. Clicking on the "…" button in the "Modifications" row will open a new window where you can adjust which modifications Byos® searches for.
3. Before filling in the custom modifications section, it is important to clear all of the preselected modifications. Click on the "…" following each row and select "Delete" until no modifications remain (Fig. 4b).
4. Now enter the following modification text in the custom modification section, and click "Ok" (Fig. 4c)—*see* **Note 1** for a brief explanation of the modification syntax:

```
mDAP / +72.0848 @ J | fixed
HexNAc(1)MurNAc_alditol(1) @ NTerm | common1
HexN(1)MurNAc_alditol(1) @ NTerm | common1

cleavage_flags=0
```

5. To enable feature finding, expand the "Feature Finder" section by clicking, then toggle the "Enable Feature Finder" option with another click (Fig. 4d).

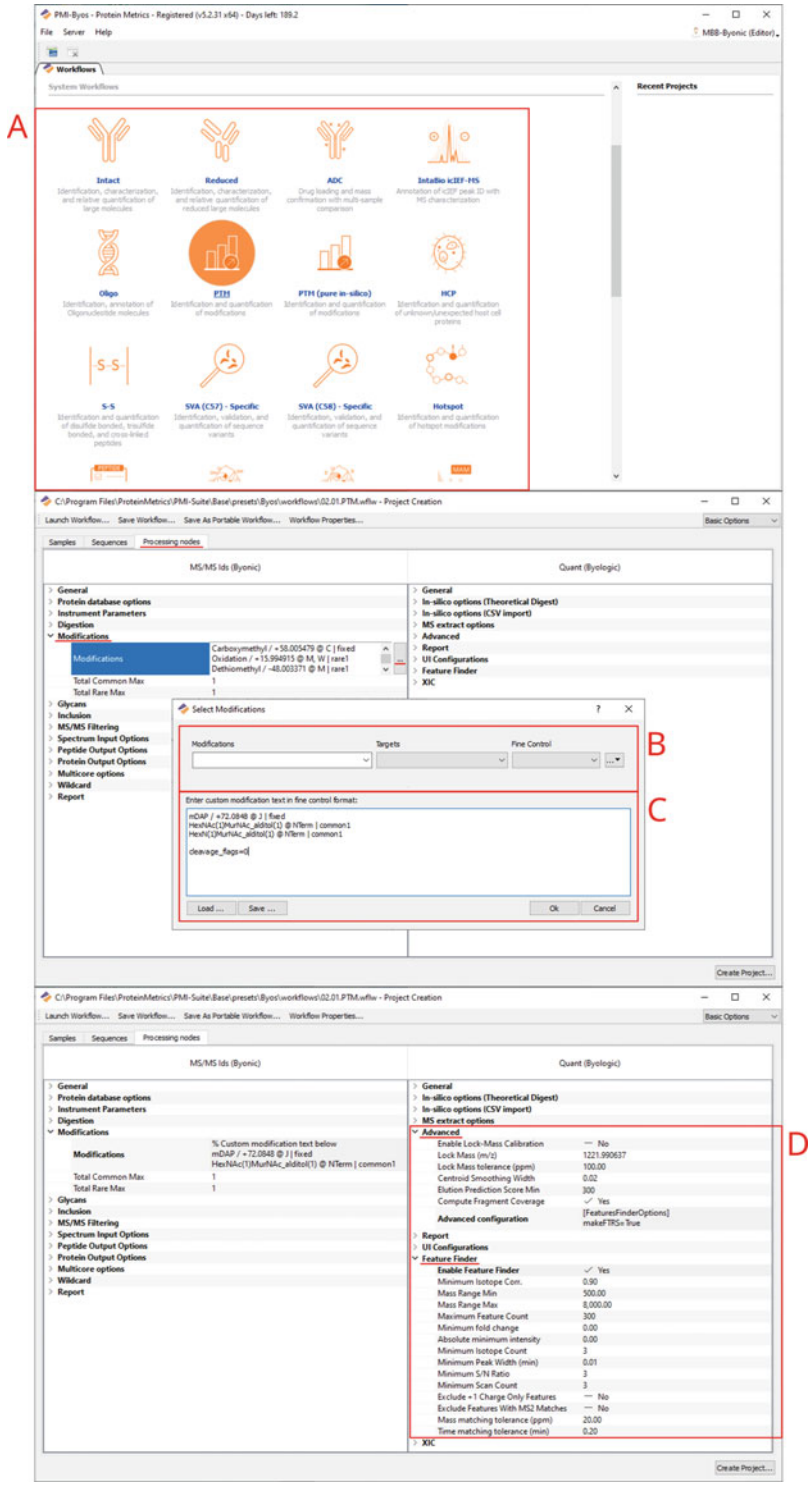


Fig. 4 Setting up a Byos workflow for data deconvolution

6. Finally, to generate .ftrs files that can be read by PGFinder, expand the "Advanced" section and add the following to the "Advanced configuration" section (Fig. 4d):

```
[FeaturesFinderOptions]
makeFTRS=True
```

7. (Optional) To save these settings for future searches, you can use the "Save Workflow" button in the menu bar to create a reloadable .wflw file for peptidoglycomics data deconvolution.
8. To actually begin the process of data deconvolution, navigate to the "Samples" tab and click "Add sample(s)..." to select your MS data files. Here, three Thermo Orbitrap .raw files containing wild-type *Escherichia coli* muropeptide data from Patel et al. [4] have been added to the samples table (Fig. 5a).
9. Then, in the "Sequences" tab, click "Add row..." and give Byos® at least one peptide to search for in your sample. Though we're not actually interested in using Byos® to identify muropeptides, deconvolution will fail if Byos® can't identify at least one structure somewhere in the data (*see* **Note 2**). For our *Escherichia coli* data, we provide the AEJ peptide, where (as the result of our custom modification settings) A is commonly modified with a GlcNAc-MurNAc disaccharide, and J represents mDAP (Fig. 5b).
10. To begin the process of deconvolution and .ftrs file generation, click "Create Project..." and provide a path that you'd like Byos to save its outputs to (Fig. 5c).

3.2 Mass Database Selection/Construction

In addition to the deconvoluted data, PGFinder separately requires a database of muropeptide structures and masses to search for. A number of mass databases for analyzing *Escherichia coli* and *Clostridium difficile* peptidoglycan have been built into the PGFinder GUI, but custom databases allow for both more refined searches and for the analysis of peptidoglycan from other bacterial species.

1. If you are analyzing peptidoglycan from *E. coli* or *C. difficile*, you can use one of PGFinder's built-in mass databases which come in three variants: Complex, Non-Redundant, and Simple. These different versions let you pick between an exhaustive, unbiased search space and a smaller, more curated search space (*see* **Note 3**).
2. To create a custom mass database, you'll need a list of structures you'd like to include and their monoisotopic masses. Using Excel, create a new sheet with two columns: "Structure" and "Monoisotopicmass" (Fig. 6a).

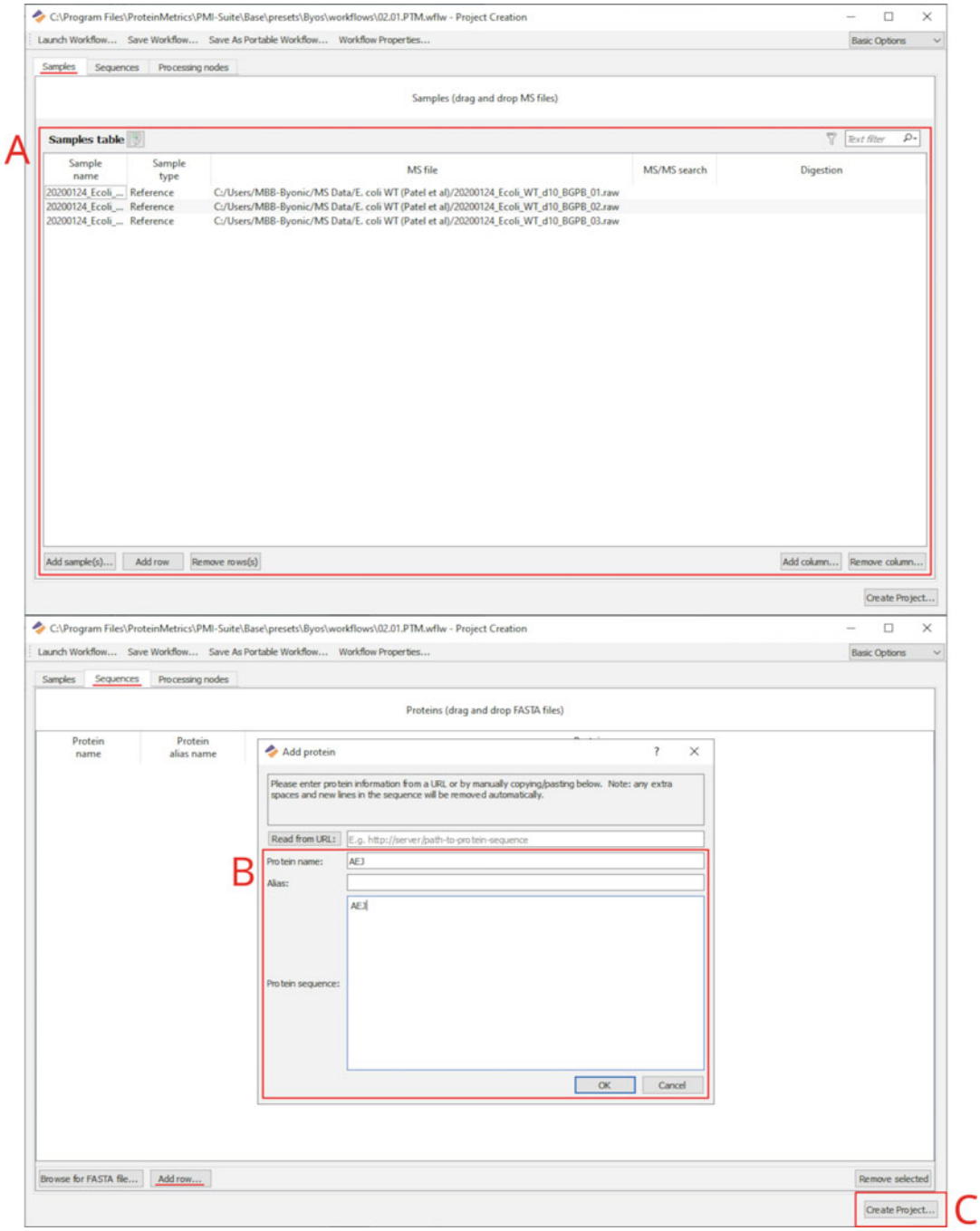


Fig. 5 Running a workflow to deconvolute multiple datasets

3. Next, copy your list of structures into the first column—importantly, you'll then need to annotate each structure with its oligomerization state. Glycan strands without peptide stems are suffixed with "|0," monomers with "|1," dimers with "|2," and so on (Fig. 6a).

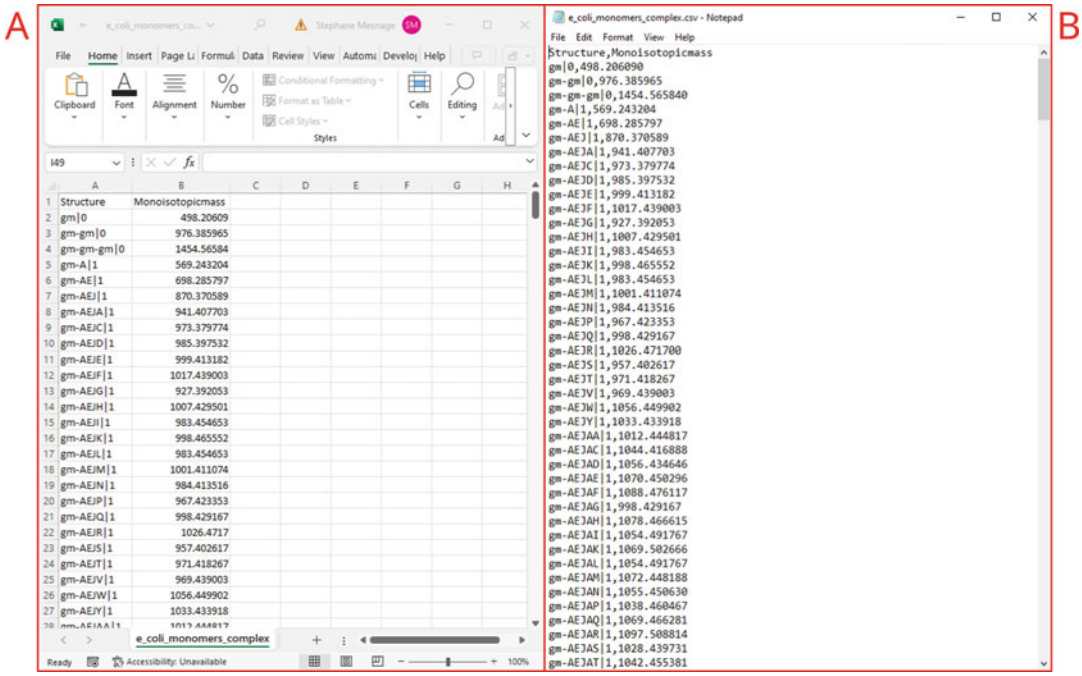


Fig. 6 Building a custom mass database and exporting it to CSV

4. Finally, copy the monoisotopic masses of each structure into the second column (Fig. 6a).
5. To export your newly constructed, custom database for use by PGFinder, save your sheet as a .csv (comma-separated values) file. You should be able to open this .csv file with a text editor like Notepad, and see something similar to Fig. 6b.

3.3 Running a PGFinder Search for Peptidoglycan Analysis

PGFinder will search your "deconvoluted" data for the monomeric muropeptides present in the provided mass database, then expand its search by building multimers and modified structures from the monomers it finds in your sample. PGFinder will then output a table of matched structures as a .csv file that can be opened using Microsoft Excel or an equivalent spreadsheet software. Here we perform a simple, one-off search, but it should be noted that more complex search strategies exist in the literature (**Note 4**).

1. MaxQuant .txt and Byos® .ftrs files can be uploaded by either dragging-and-dropping files into the "Deconvoluted Datasets" dropzone or by clicking on the drop zone to browse for data files. Multiple datasets can be selected for batch processing using the same mass database, modifications, and search options (Fig. 7a).
2. Here, we have elected to use the built-in "Simple" mass database for *E. coli* (Fig. 7b),

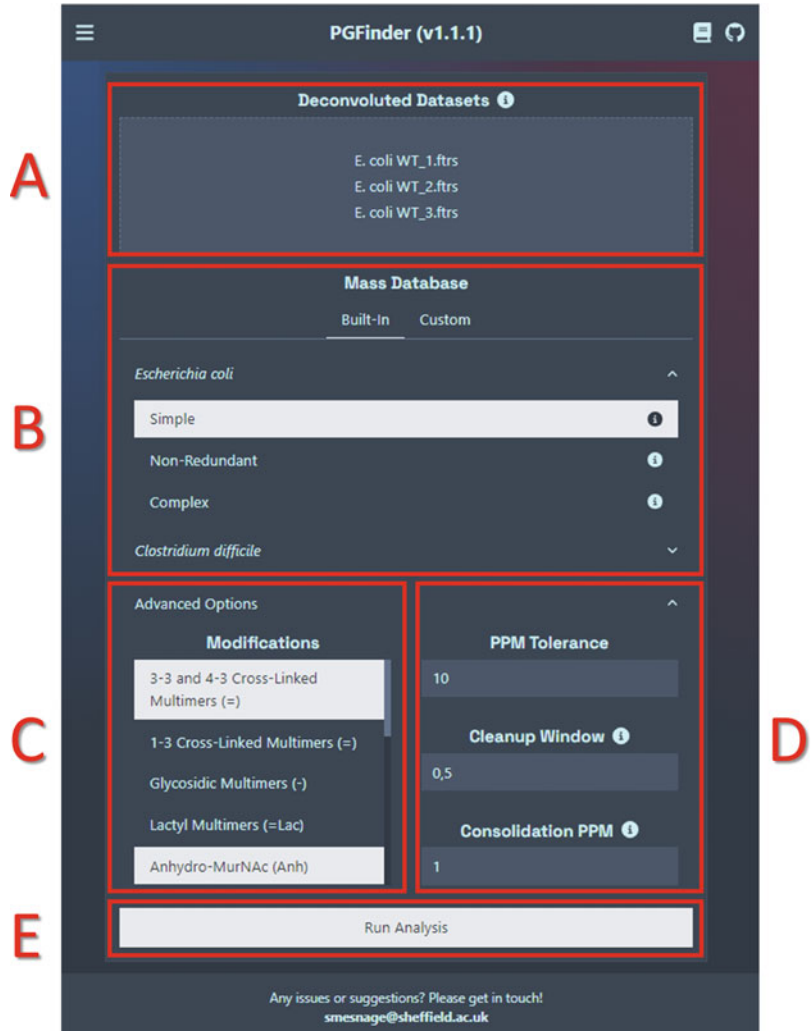
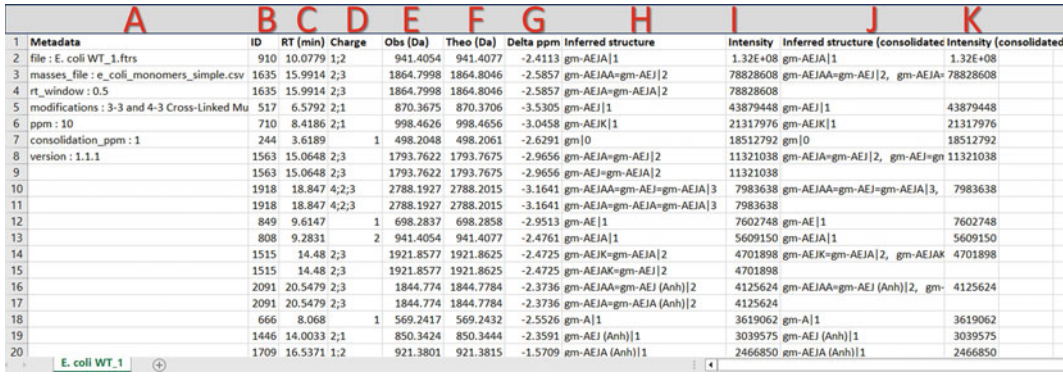


Fig. 7 The PGFinder WebUI

3. To specify which multimers and modifications you'd like PGFinder to construct and look for, first click on "Advanced Options," then scroll through the list of "Modifications" and select the ones you'd like to search for—here we've selected "3-3 and 4-3 Cross-Linked Multimers (=)," "Anhydro-MurNAc (Anh)," and "Deacetylation (-Ac)" (Fig. 7c).
4. (Optional) Several more search parameters are available to adjust under "Advanced Options," including "PPM Tolerance," "Cleanup Window," and "Consolidation PPM" (*see* **Note 5** for more detail). In Fig. 7d, all options have been left at their defaults.



	A	B	C	D	E	F	G	H	I	J	K
1	Metadata	ID	RT (min)	Charge	Obs (Da)	Theo (Da)	Delta ppm	Inferred structure	Intensity	Inferred structure (consolidatec	Intensity (consolidated)
2	file : E. coli WT_1.fters	910	10.0779	1;2	941.4054	941.4077	-2.4113	gm-AEJA|1	1.32E+08	gm-AEJA|1	1.32E+08
3	masses_file : e_coli_monomers_simple.csv	1635	15.9914	2;3	1864.7998	1864.8046	-2.5857	gm-AEJAA=gm-AEJ|2	78828608	gm-AEJAA=gm-AEJ|2, gm-AEJA=	78828608
4	rt_window : 0.5	1635	15.9914	2;3	1864.7998	1864.8046	-2.5857	gm-AEJA=gm-AEJA|2	78828608		
5	modifications : 3-3 and 4-3 Cross-Linked Mu	517	6.5792	2;1	870.3675	870.3706	-3.5305	gm-AEJ|1	43879448	gm-AEJ|1	43879448
6	ppm : 10	710	8.4186	2;1	998.4626	998.4656	-3.0458	gm-AEJK|1	21317976	gm-AEJK|1	21317976
7	consolidation_ppm : 1	244	3.6189	1	498.2048	498.2061	-2.6291	gm|0	18512792	gm|0	18512792
8	version : 1.1.1	1563	15.0648	2;3	1793.7622	1793.7675	-2.9656	gm-AEJA=gm-AEJ|2	11321038	gm-AEJA=gm-AEJ|2, gm-AEJ=gm	11321038
9		1563	15.0648	2;3	1793.7622	1793.7675	-2.9656	gm-AEJ=gm-AEJA|2	11321038		
10		1918	18.847	4;2;3	2788.1927	2788.2015	-3.1641	gm-AEJAA=gm-AEJ=gm-AEJA|3	7983638	gm-AEJAA=gm-AEJ=gm-AEJA|3,	7983638
11		1918	18.847	4;2;3	2788.1927	2788.2015	-3.1641	gm-AEJA=gm-AEJA=gm-AEJA|3	7983638		
12		849	9.6147	1	698.2837	698.2858	-2.9513	gm-AE|1	7602748	gm-AE|1	7602748
13		808	9.2831	2	941.4054	941.4077	-2.4761	gm-AEJA|1	5609150	gm-AEJA|1	5609150
14		1515	14.48	2;3	1921.8577	1921.8625	-2.4725	gm-AEJK=gm-AEJA|2	4701898	gm-AEJK=gm-AEJA|2, gm-AEJAK	4701898
15		1515	14.48	2;3	1921.8577	1921.8625	-2.4725	gm-AEJAK=gm-AEJ|2	4701898		
16		2091	20.5479	2;3	1844.774	1844.7784	-2.3736	gm-AEJAA=gm-AEJ (Anh)|2	4125624	gm-AEJAA=gm-AEJ (Anh)|2, gm-	4125624
17		2091	20.5479	2;3	1844.774	1844.7784	-2.3736	gm-AEJA=gm-AEJA (Anh)|2	4125624		
18		666	8.068	1	569.2417	569.2432	-2.5526	gm-A|1	3619062	gm-A|1	3619062
19		1446	14.0033	2;1	850.3424	850.3444	-2.3591	gm-AEJ (Anh)|1	3039575	gm-AEJ (Anh)|1	3039575
20		1709	16.5371	1;2	921.3801	921.3815	-1.5709	gm-AEJA (Anh)|1	2466850	gm-AEJA (Anh)|1	2466850

E. coli WT_1

Fig. 8 A (truncated) output from PGFinder

5. Finally, clicking "Run Analysis" will set PGFinder running on all of the uploaded data files. Upon completion, PGFinder will trigger the download of one .csv file for every initially supplied data file (Fig. 7e).

3.4 Analyzing PGFinder's Results

3.4.1 PGFinder's Output

The output columns can be divided into three groups: (A) the metadata column, (B-I) the "long format" columns, and (J-K) the consolidated matches (Fig. 8). PGFinder lists all of the matched masses at the top of each .csv file (sorted by intensity), then lists any unmatched masses below that (not shown).

A. The metadata column is populated with information describing the search that was performed to generate these results. It includes the name of the deconvoluted data file, the name of the mass database, the enabled modifications, any other search parameters, and the version of PGFinder used.

B. The ID column displays the unique ID of each feature generated during feature finding/data deconvolution. Note that some rows contain duplicate IDs—this indicates that two or more theoretical structures matched the same observed monoisotopic mass.

C. The RT (min) column simply records the retention time (in minutes) at which each mass was eluted from the chromatographic column.

D. The Charge column records the charge(s) at which each mass was found. When deconvoluting data with MaxQuant, this will be a single number (each row represents one ion), but when deconvoluting using Byos®, this is often a list of several charges separated with a semicolon (;)—in this case, each row represents one monoisotopic mass.

E. The Obs (Da) column records the monoisotopic mass (in daltons) observed by the mass spectrometer for each matched and unmatched ion.

F. The Theo (Da) column records the theoretical monoisotopic mass (in daltons, from the user-provided mass database) for each matched ion.

G. The Δppm column quantifies the distance between each observed and theoretical mass in parts per million.

H. The Inferred structure column shows which theoretical structure (from the user-provided mass database) matched each ion.

I. The Intensity column records the integrated XIC value for each ion as reported by either MaxQuant or Byos and can be used as a reliable metric for relative ion quantification.

J. As previously mentioned, sometimes a single observed mass can match several theoretical structures—in columns B-I, this results in two or more rows that all correspond to the same ion. To prevent the intensities of these duplicate rows from skewing downstream quantification, PGFinder selects the structure that most closely matches the observed mass and reports that in the Inferred structure (consolidated) column. If the absolute Δppm of several potential masses are less than the consolidation ppm apart, then all of those matches are included (separated by a comma and three spaces — ",") in this column. In this column, each row corresponds to a unique ion ID.

K. The Intensity (consolidated) column simply copies the intensity of each unique ion into a new column, avoiding the double counting of any ion intensities during downstream analysis.

3.4.2 Consolidating Individual Search Outputs

While much of the information in PGFinder's output is immediately useful (the list of structures and retention times, for example, enables chromatogram annotation), further processing is needed to arrive at a publication-ready table or figure.

1. Before processing the results from PGFinder, the .csv files corresponding to each biological replicate need to be imported as an individual tab into a single Excel workbook. To do this, navigate to the "Data" tab, click "From Text/CSV" in the "Get & Transform Data" section, and select the .csv file created by PGFinder for your first biological replicate. Verify that the correct file has been loaded in the preview screen, then click "Load" to load your data into a new sheet—repeat this process for the other two biological replicates (Fig. 9a).
2. Sometimes the same structure can match ions present at several retention times, so to generate a list of unique structures and their pooled intensities, we make use of Excel's consolidation functionality. Select the area where you would like the consolidated data to be written (cell M1 was selected in Fig. 9b), then in the "Data" tab, click "Consolidate" in the "Data Tools" section. In the pop-up, select "Sum" as the "Function" and

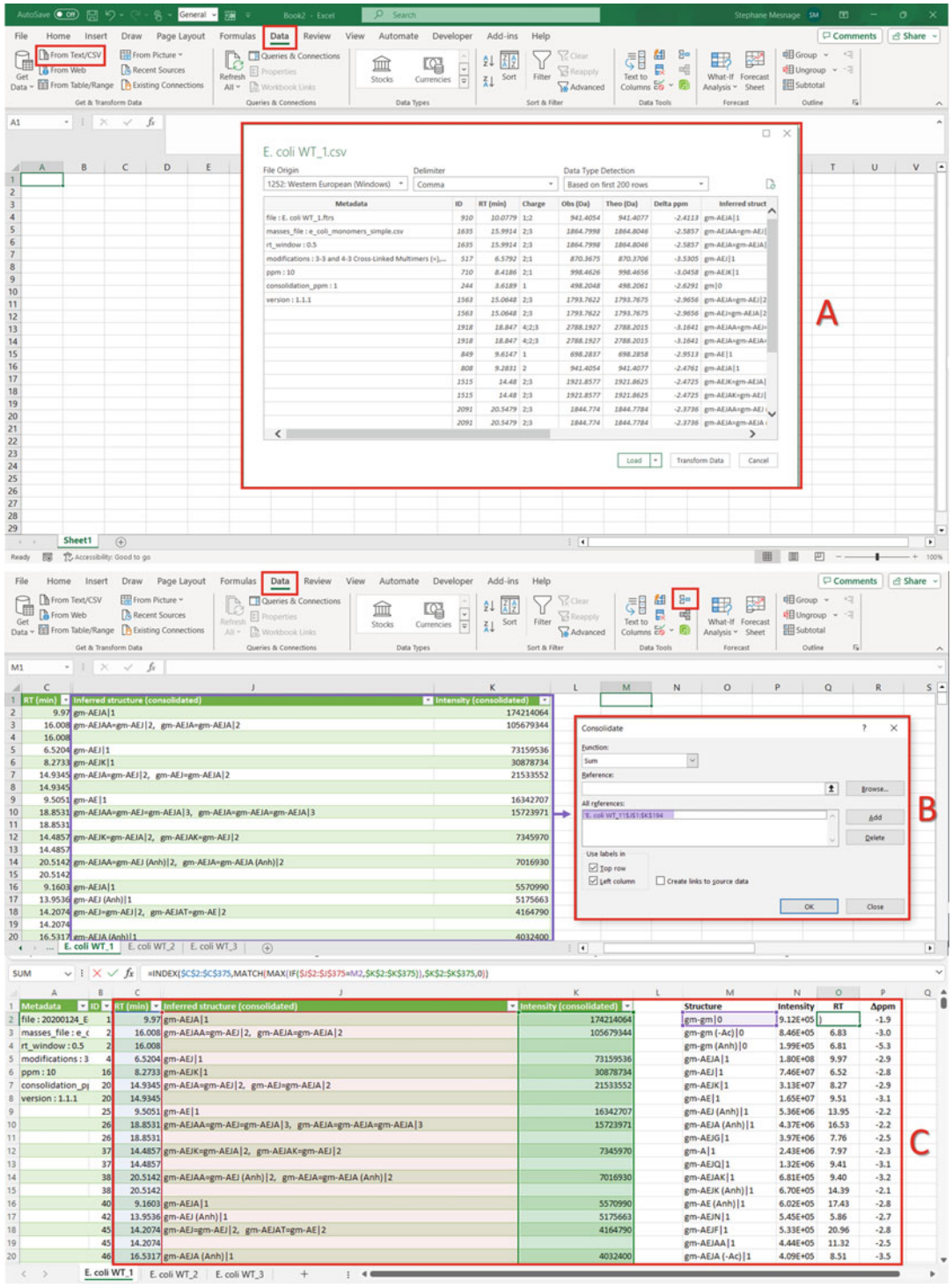


Fig. 9 Completing the consolidation of a single dataset in Excel

click the up-arrow button under "Reference" to select the "Inferred structure (consolidated)" and "Intensity (consolidated)" columns of PGFinder's output. Finally, click "Add," ensure that both "Top row" and "Left column" are selected under the "Use labels in" section, and click "OK" to consolidate (Fig. 9b).

3. The consolidation output can then be briefly cleaned up by naming the first column "Structure," renaming the second "Intensity," and deleting the blank consolidated row. After this process, the first two columns should resemble columns M and N from Fig. 9c.
4. At this stage, it is vitally important to inspect all of the mass coincidences and ambiguities reported by PGFinder. Often, prior knowledge can be used to discount certain structures that are identified as the result of common mass coincidences, and many other cases can have the correct structure inferred from the abundance of acceptors, but ultimately a manual inspection of the MS/MS data may be needed to discriminate between isomeric structures—notably when discriminating between 3-3 and 4-3 cross-links. *The details of this process, however, are out of scope for this chapter, so all ambiguities have been resolved here by simply selecting the first structure listed in the "Inferred structure (consolidated)" column.*
5. To finish the consolidation process, the Excel formula found in Fig. 9c — `=INDEX(<RT column>,MATCH(MAX(IF(<Consolidated structure column>=<Current structure>, <Intensity column>)),<Intensity column>,0))` — is used to look up the retention time at which each consolidated structure was found at its maximum intensity (*see* **Note 6** for a more detailed explanation). This same formula with some minor adjustments can then be used to copy over the Δppm of each consolidated structure, resulting in the final table shown in Fig. 9c—repeat this entire consolidation process for each of the other two biological replicates.

3.4.3 Consolidating Biological Replicates

1. After each biological replicate has been individually consolidated, they can be averaged together to obtain a mean intensity, retention time, and Δppm as well as a standard deviation for each. To do this, create a new sheet for the consolidated data, select the area you'd like to write the consolidated data to (A2 in Fig. 10), and once again click "Consolidate" under the "Data Tools" section to open the consolidation window. This time select "Average" for the function and add three references—one covering the previously consolidated "Structure," "Intensity," "RT," and "Δppm" columns for each biological replicate. Click "OK" to populate the new sheet with averaged values from all three biological replicates (Fig. 10a).

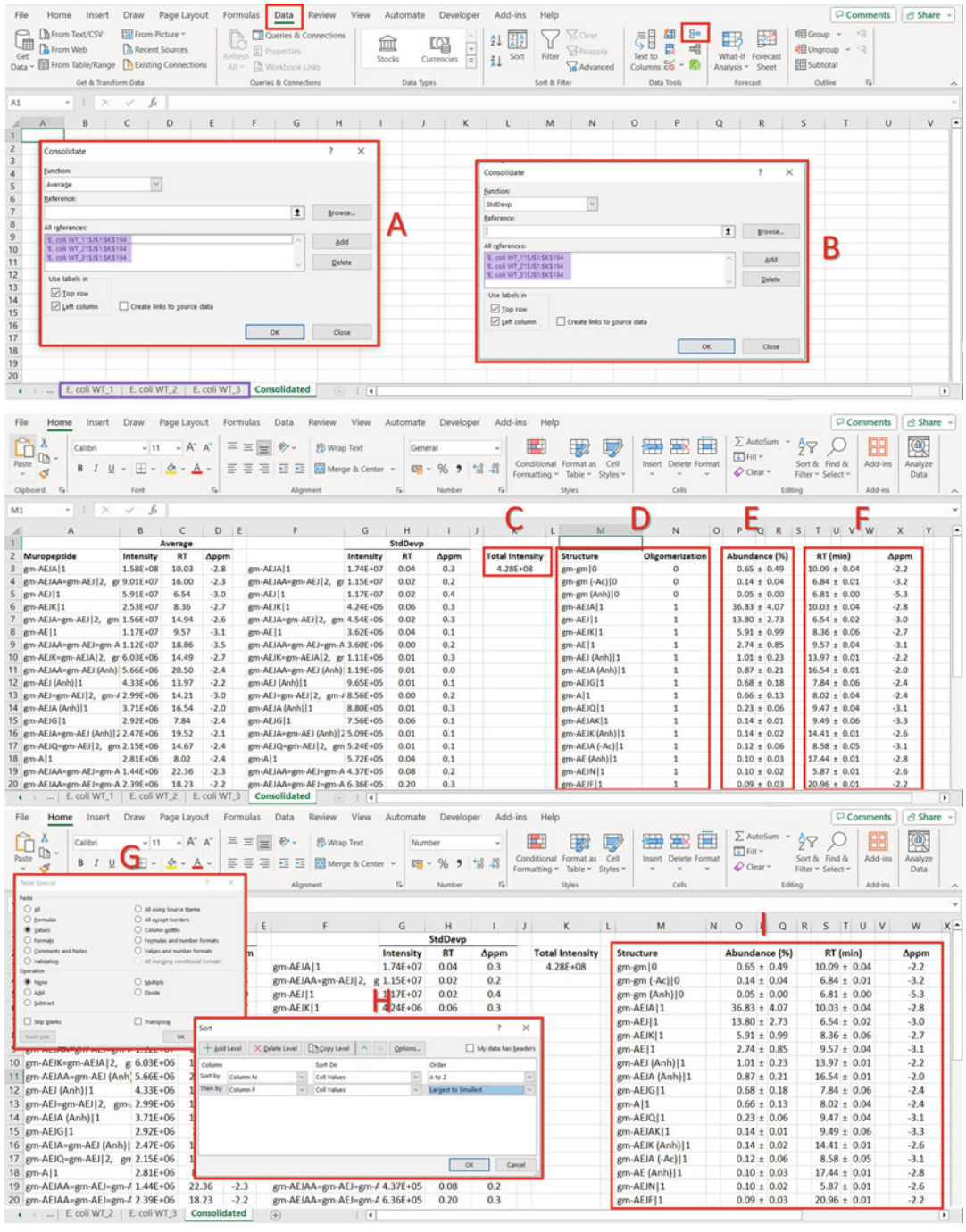


Fig. 10 Consolidating biological replicates into a final muropeptide table in Excel

2. Next, select a new location for the consolidated standard deviations (F2 in Fig. 10), click "Consolidate," and select "StdDevp" as the function. Use the same three references as in **Step 1**, and click "OK" (Fig. 10b).
3. In the final table of muropeptides, it's helpful to convert raw intensities into relative abundances (%)—to do so, we first need to sum all of the average intensities by creating a new "Total Intensity" cell using Excel's `=SUM(...)` function (Fig. 10c).
4. To start building the final table of muropeptides, copy over all of the consolidated structures, then create a temporary "Oligomerization" column for storing the oligomerization state of each structure, which will be used to later sort structures into monomers, dimers, and trimers. To extract the last digit of each structure, you can use the `=RIGHT(...,1)` Excel function — for N3 in Fig. 10d, you'd write `=RIGHT(M3,1)` (Fig. 10d).
5. Next, relative abundances can be calculated by dividing each structure's intensity by the "Total Intensity" calculated in **step 3**, then multiplying by 100 (e.g., for cell P3 in Fig. 10, `=B3/$K$3*100`). The standard deviation of each structure's intensity can be converted into relative abundance in the same way (Fig. 10e).
6. The retention times and Δppms can be copied over into the final table alongside their respective standard deviations (often omitted for Δppm) without the need any further rescaling or transformation (Fig. 10f).
7. To sort the muropeptide table by abundance, the formula cells first need to be replaced by their values. To do this, select the whole muropeptide table (columns M–W in Fig. 10) and copy the cells. Next, with the same area selected, right-click and choose "Paste Special…." then select "Values" under the "Paste" section and click "OK" (Fig. 10g).
8. Next, with the whole table selected, right-click again and select "Sort," then "Custom Sort…." In the dialog, sort first by the "Oligomerization" column (column N in Fig. 10d), select "Cell Values" for "Sort On," then choose "A to Z" for the "Order." To then sort each category by abundance, click "Add Level," then sort by the "Abundance %" column (column P in Fig. 10e), choose "Cell Values," then choose "Largest to Smallest." Clicking "OK" will sort the muropeptide table by monomers, dimers, and trimers, then abundance (Fig. 10h).
9. Finally, you can delete the temporary "Oligomerization" column to arrive at the final table of consolidated and sorted muropeptides seen in Fig. 10i.

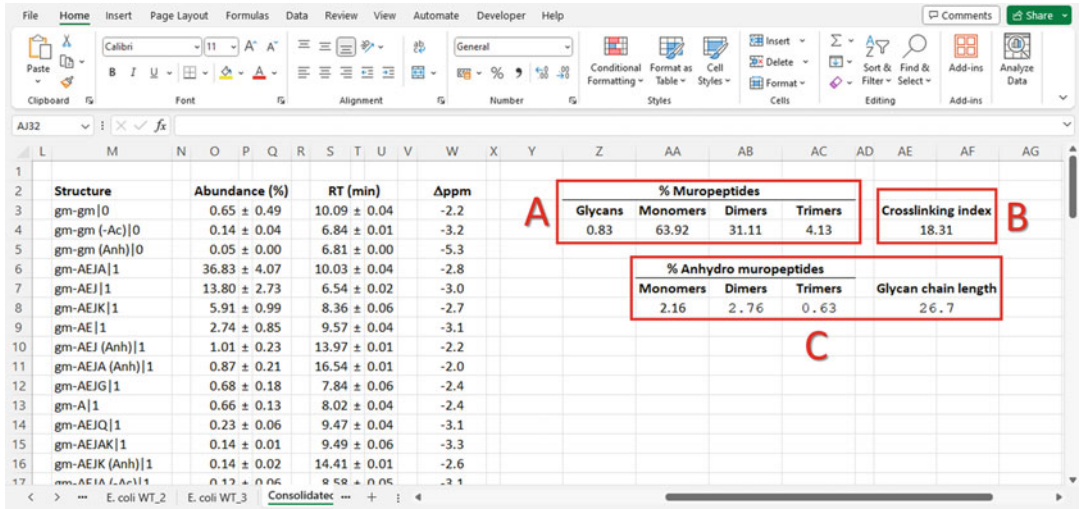


Fig. 11 Calculating a number of commonly reported metrics

3.4.4 Calculating Commonly Reported Metrics

1. To calculate the distribution of monomers, dimers, and trimers, the `=SUMIF(...)` function from Excel can be used. Looking at the number after the "|" in each structure, we can find monomers ("|1"), dimers ("|2"), and trimers ("|3"). Giving SUMIF the column of structures, a pattern to match ("*|1" for monomers) and then the column to sum, we arrive at the final formula for monomer abundance in Fig. 11a: `=SUMIF(M3:M84,"*|1",O3:O84)`. To adapt this formula for dimers and trimers, simply swap out the "*|1" for "*|2" and "*|3" (Fig. 11a).
2. From this distribution of monomers, dimers, and trimers, the cross-linking index (the percentage of mDAP residues involved in cross-linking) can be calculated according to the following formula [1]: ½ × % dimers + ⅔ × % trimers. The result of translating this equation into an Excel formula can be seen in Fig. 11b.
3. Finally, glycan chain length is calculated from the abundance of anhydroMurNAc containing monomers, dimers, and trimers. To calculate these abundances, copy the `SUMIF` formula from **step 1** and add "(Anh)" to the search pattern. To locate anhydro-monomers, for example, adjust the formula in Fig. 11a to read: `=SUMIF(M3:M84,"*(Anh)|1",O3:O84)`. Then, to calculate glycan chain length, use the following formula [1]: 100/(% anh-monomers + ½ × % anh-dimers + ⅓ × anh-trimers). Calculating this in Excel, using the newly calculated abundances of the anyhydroMurNAc containing structures, gives the glycan chain length found in Fig. 11c

4 Notes

1. Focusing on the first three (3) lines of modification text, each line follows the pattern: `<modification> @ <location> | <prevalence>`. The first line contains a custom modification representing meso-diaminopimelic acid called "mDAP" which adds 72.0848 daltons to J's Byos®-defined mass of 100 daltons, bringing the total to 172.0848 daltons for mDAP. The "fixed" marker means that *every* J residue will contain this modification. The next two lines define modifications using Byos®'s built-in glycans. The first line represents a GlcNAc-MurNAc disaccharide (with GlcNAc being called HexNAc, and MurNAc being in its reduced, alditol form, and with the (1) markers meaning one of each sugar is present). The location for these glycans is the N-terminal of the peptide, and we expect these to be common modifications present in a single copy (1) on each peptide. The third line simply represents a deacetylated version of the disaccharide defined in line two. Finally, the `cleavage_flags=0` option at the end prevents Byos® from automatically generating peptides that have their initial residue cleaved off (as is common in proteomics datasets, but not commonly the case for peptidoglycan).
2. Since Byos® will fail to deconvolute any data if it can't find the peptide you supply here, you'll need to be sure to pick one that will actually be present in your sample! While many diderms (gram-negative bacteria) will contain an AEJ stem peptide, many monoderms (gram-positive bacteria) will contain a different tripeptide (e.g., *Staphylococcus aureus* builds its PG from AQK building blocks). Consequently, picking the right peptide during this step requires some a priori knowledge regarding the composition of your sample or a bit of trial and error.
3. If you'd like to know more about the contents of each built-in database, hover your cursor over the ⓘ and a popup should appear with more information. If you'd like to dig even deeper, you can see the exact contents of each database on GitHub: https://github.com/Mesnage-Org/pgfinder/tree/master/lib/pgfinder/masses
4. For recently published examples of more complex search strategies, see the supplementary figures of both [8] and [9].
5. "PPM Tolerance" corresponds to the maximum PPM difference allowed between observed and theoretical monoisotopic masses during the matching process. "Cleanup Window" defines the retention time window within which the intensities of an ion's salt adducts and in-source decay products (loss of GlcNAc) are consolidated. If multiple structures are possible for a given observed mass, PGFinder will select the structure

with the lowest absolute PPM to place in the "Inferred structure (consolidated)" column, but if two matches differ from the observed mass by less than the "Consolidation PPM," then both are included in PGFinder's output. Setting the "Consolidation PPM" equal to the "PPM Tolerance" will effectively disable this "closest match" picking and will retain all possible matches in the consolidated structures column.

6. This Excel formula is best understood when looked at from the inside-out. Starting with the IF, we test if each structure in the "Inferred structure (consolidated)" column matches the structure for this row and, if it does, return the corresponding "Intensity (consolidated)" for that matching row. This results in a list of intensities for each of the times this row's structure appears in PGFinder's output. To find the largest intensity of those in this list, Excel's `MAX` function is used. Once the maximum intensity is found, the `MATCH` function is used to find which row number that maximum value came from (the 0 at the end tells `MATCH` to look for an exactly matching value). Once we know which row contains the maximum intensity for our structure, we can use `INDEX` to look up the RT or Δppm value of that row. To switch between fetching RT's and Δppm's, you only need to change which column is supplied to `INDEX`

Acknowledgments

BJR is the recipient of a NERC PhD studentship (NERC ACCE NE/S00713X/1). The work on PGFinder in S. Mesnage's lab was supported by grants from the MRC (MR/S009272/1) and the BBSRC (BB/W013800/1).

References

1. Glauner B (1988) Separation and quantification of muropeptides with high-performance liquid chromatography. Anal Biochem 172(2): 451–464. https://doi.org/10.1016/0003-2697(88)90468-x
2. Anderson EM, Sychantha D, Brewer D, Clarke AJ, Geddes-McAlister J, Khursigara CM (2020) Peptidoglycomics reveals compositional changes in peptidoglycan between biofilm- and planktonic-derived Pseudomonas aeruginosa. J Biol Chem 295(2):504–516. https://doi.org/10.1074/jbc.RA119.010505
3. Hsu Y-C, Su P-R, Huang L-J, Cheng K-Y, Chen C, Hsu C-C (2023) High-throughput Automated Muropeptide Analysis (HAMA) reveals peptidoglycan composition of gut microbial cell walls. 17. https://doi.org/10.1101/2023.04.17.537164
4. Patel AV et al (2021) PGFinder, a novel analysis pipeline for the consistent, reproducible, and high-resolution structural analysis of bacterial peptidoglycans. elife 10:e70597. https://doi.org/10.7554/eLife.70597
5. Deutsch EW (2010) Mass spectrometer output file format mzML. In: Hubbard SJ, Jones AR (eds) Proteome bioinformatics, Methods in molecular biology™. Humana Press, Totowa, pp 319–331. https://doi.org/10.1007/978-1-60761-444-9_22
6. Chambers MC et al (2012) A cross-platform toolkit for mass spectrometry and proteomics.

Nat Biotechnol 30(10):Art. no. 10. https://doi.org/10.1038/nbt.2377

7. Cox J, Mann M (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol 26(12): Art. no. 12. https://doi.org/10.1038/nbt.1511
8. Alamán-Zárate MG et al (2024) Unusual 1-3 peptidoglycan cross-links in Acetobacteraceae are made by L,D-transpeptidases with a catalytic domain distantly related to YkuD domains. J Biol Chem 300(1). https://doi.org/10.1016/j.jbc.2023.105494
9. Galley NF et al (2023) Clostridioides difficile canonical L,D-transpeptidases catalyse a novel type of peptidoglycan cross-links and are not required for beta-lactam resistance. J Biol Chem 300(1). https://doi.org/10.1016/j.jbc.2023.105529