

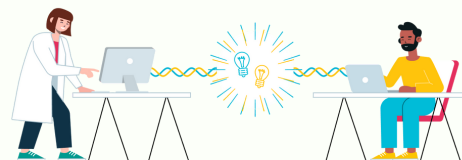
Technical Note

Transforming data to knowledge using
Mass Dynamics™ Generic Format



Why we built it

- Mass Spectrometry-based Proteomics (MSP) analysis is a powerful and flexible approach. One of the major challenges in its use is the need to adopt multiple tools to accurately evaluate data quality, identify, measure protein abundance and finally translate this analysis into useful biological information.
- Irrespective of how MSP data is acquired and processed, a generic format will make it easier to leverage the powerful features of [Mass Dynamics 1.0](#) [1] to generate useful insights, such as:
 - automated quality control reports
 - statistical evaluations
 - biological knowledge insight generation using enrichment analysis or over representation analysis on your differentially expressed protein data
 - evaluating the data quality and results faster
 - sharing knowledge directly with your collaborators
 - giving collaborators direct access to explore datasets independently and interactively



What does this mean for you?

- No matter what software you use to pre-process data, you can adopt Mass Dynamics to seamlessly generate key insights and share results with collaborators
- Give your proteomics data a single home, bringing together the generated knowledge and expertise to make informed decisions faster than ever before
- And ultimately, avoid:
 - Spending time and energy patching together siloed tools and software packages
 - Second-guessing the quality of experiment results
 - Working inefficiently with your multidisciplinary life science team (say 'goodbye' zip files and ftp servers, 'hello' web-based collaboration!)
 - Manually reviewing online knowledge bases and potentially creating bias due to cherry-picking interesting insights
 - Being the bottleneck in generating reports and sharing insights with stakeholders
 - Analysis delays due to reliance on experts to navigate the data

Background

Evaluation of complex samples using liquid chromatography mass spectrometry (LC-MS) generates information rich data that is difficult to interpret and align back to biological processes. Typical processing of the data results in a list of identified proteins and quantitative abundance changes between conditions.

Many software packages which process LC-MS data, such as MaxQuant or Proteome Discoverer (Thermo Fisher), produce similar outputs of summarized protein intensities. In such an output, each row corresponds to the intensities of a protein detected across samples. These signals are often already aggregated across fractions and for certain experiment designs, such as tandem mass tag experiments (TMT), may correspond to specific channels.

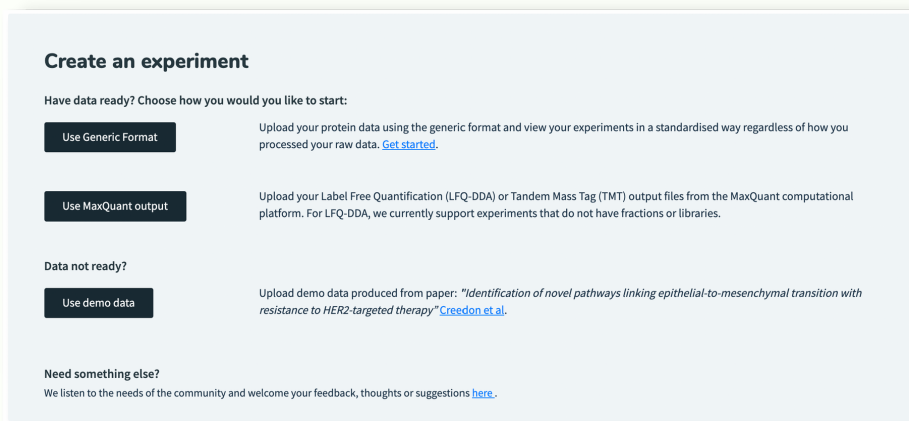


With 'simplicity' at our core, we have expanded our existing analysis pipeline in which data can be uploaded. This enables users to process any summarized protein intensity data. This means no matter what software you use to pre-process raw MSP data, you can transform this to knowledge seamlessly using MassDynamics 1.0.

Analyse MS Data using a generic format

Step 1 - Sign up and start

Simply **sign up**, select 'Create an Experiment' and choose the 'Use Generic Format' option and follow the instructions.



Create an experiment

Have data ready? Choose how you would you like to start:

- Use Generic Format** Upload your protein data using the generic format and view your experiments in a standardised way regardless of how you processed your raw data. [Get started.](#)
- Use MaxQuant output** Upload your Label Free Quantification (LFQ-DDA) or Tandem Mass Tag (TMT) output files from the MaxQuant computational platform. For LFQ-DDA, we currently support experiments that do not have fractions or libraries.

Data not ready?

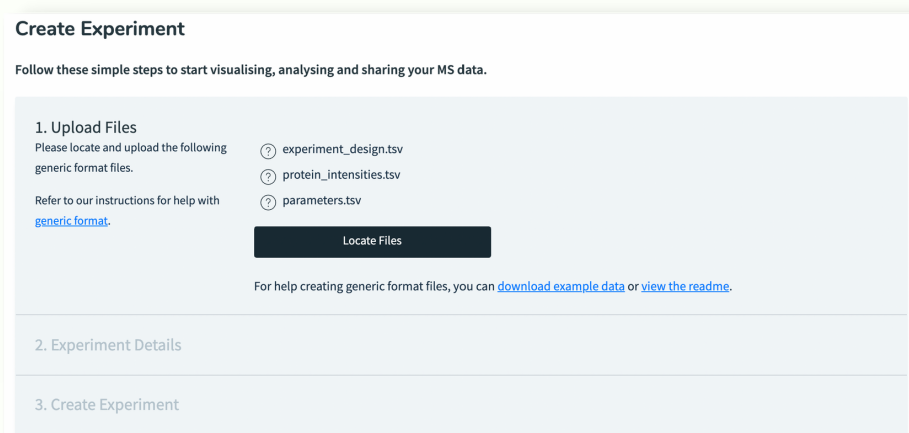
- Use demo data** Upload demo data produced from paper: "Identification of novel pathways linking epithelial-to-mesenchymal transition with resistance to HER2-targeted therapy" [Creedon et al.](#)

Need something else?
We listen to the needs of the community and welcome your feedback, thoughts or suggestions [here.](#)

Step 2 - Locate MS data files

Mass Dynamics' generic format requires three (3) tab separated files:

1. Experiment design file
2. Protein intensity file
3. Parameters file



Create Experiment

Follow these simple steps to start visualising, analysing and sharing your MS data.

- 1. Upload Files**
Please locate and upload the following generic format files.
Refer to our instructions for help with [generic format.](#)
 - experiment_design.tsv
 - protein_intensities.tsv
 - parameters.tsv

Locate Files

For help creating generic format files, you can [download example data](#) or [view the readme.](#)
2. Experiment Details
3. Create Experiment

Step 2 (cont.)

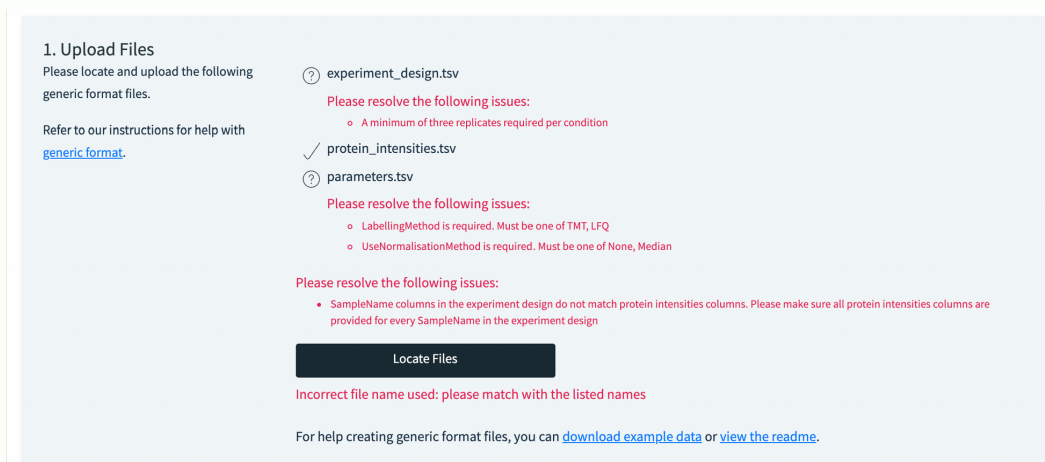
The table below summarises the specific requirements for each file:

FILE	REQUIRED COLUMN HEADINGS	NOTES
Protein Intensities		Each row corresponds to detected intensities for one protein in all samples
	ProteinId	Uniprot accession number (one per row)
	Samples intensities names <i>(any name you use for samples analysed)</i>	<ul style="list-style-type: none"> This needs to correspond to the entries of the SampleName column in the Experiment Design file Intensities should contain numeric, not logged data. Data can be missing and 0's will be treated as missing values.
		GeneName (GeneSymbols) and gene Descriptions as columns are optional
Experiment Design	SampleName	Each entry needs to correspond to one column name in the protein intensities table.
	Condition	Indicates the groups which will be used for the differential expression and enrichment analyses.
Parameters		Each row should contain a parameter and its selection separated by a tab.
	Species	Choose one of "Human", "Mouse", "Yeast" or "Other". This determines the gene set libraries used in the Enrichment feature.
	UseNormalisationMethod	Select "None" for no normalization or "Median" if you wish to perform this as part of the upload. If 'Median' is selected, the median of each logged intensity column is subtracted from the columns prior to imputation/statistical analysis.
	LabellingMethod	Define whether the experiment is LFQ or TMT

Table 1: Minimum required information to upload generic format and access Mass Dynamics 1.0 services.

Step 3 - Confirm format of files

After you complete locating all files, Mass Dynamics will automatically validate the contents of each file to ensure correct formats and provide feedback if needed. An example is in Figure 1 below.



The screenshot shows a web interface for uploading files. It lists three files: 'experiment_design.tsv', 'protein_intensities.tsv', and 'parameters.tsv'. 'experiment_design.tsv' and 'parameters.tsv' have red question marks and associated error messages. 'protein_intensities.tsv' has a green checkmark. A 'Locate Files' button is visible, along with a red error message: 'Incorrect file name used: please match with the listed names'. At the bottom, there is a link for help: 'For help creating generic format files, you can [download example data](#) or [view the readme](#).'

1. Upload Files
Please locate and upload the following generic format files.

Refer to our instructions for help with [generic format](#).

experiment_design.tsv
Please resolve the following issues:
• A minimum of three replicates required per condition

protein_intensities.tsv

parameters.tsv
Please resolve the following issues:
• LabellingMethod is required. Must be one of TMT, LFQ
• UseNormalisationMethod is required. Must be one of None, Median

Please resolve the following issues:
• SampleName columns in the experiment design do not match protein intensities columns. Please make sure all protein intensities columns are provided for every SampleName in the experiment design

Locate Files

Incorrect file name used: please match with the listed names

For help creating generic format files, you can [download example data](#) or [view the readme](#).

Figure 1. Example of feedback provided after automatic file validation.

Step 4 - Finish creating experiment

Follow steps to complete Experiment details and select Create Experiment.

Step 5 - Sit back while we process your files

Mass Dynamics will upload all files and begin processing data using the Mass Dynamics 1.0 processing workflow, which includes:

- **Log-transformation** of raw protein intensities (Note: If normalization is requested, the median intensity of each sample will be subtracted from the logged data)
- Detection of **missing values** will be imputed using Missing Not At Random (MNAR) imputation
- Setup of **pairwise comparisons** (Note: Linear models are fitted for each pairwise comparison using limma [2]). For each comparison, only proteins with at least 50% available measurements are considered
- Calculation of **Enrichment** results for Gene Ontologies (GO) and Reactome [3,4]
- Creation of a **QC Report** to address experimental health, data quality, and the effects of imputation and normalization (if selected)

Step 6 - Review the Quality Control (QC) Report

Data uploaded using the generic format is easily assessed using an automatically generated QC Report, which includes:

- PCA plots using all and differentially expressed proteins to confirm known clustering of the samples or to identify unusual patterns
- Quantitative % CV distributions across samples
- Intensity distribution of features before and after normalization

All information can be downloaded and shared as a complete QC report.

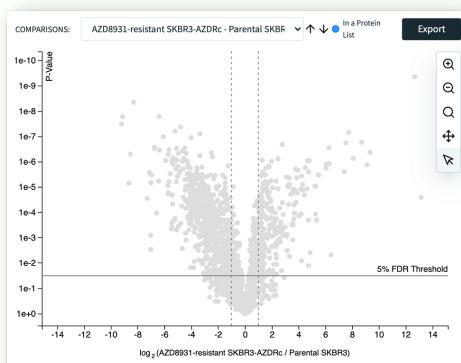
Please refer to the [Quality Control Report video](#) for more details on how to use the report.



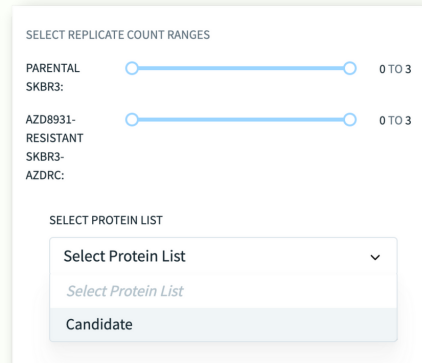
Step 7 - Happy analysing!

With your data setup and ready for analysis, you can now make sense of your MSP data and get closer to your next moment of discovery.

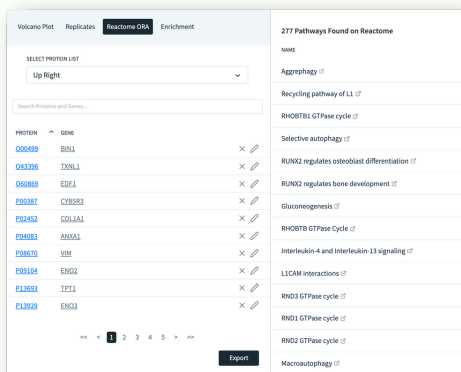
Here are some example features. Please refer to the Enrichment [video](#) and [App Note](#) for more details.



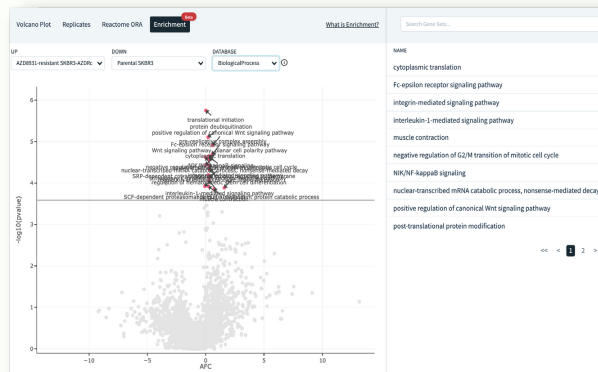
Above: Pairwise comparison



Above: Replicate analysis

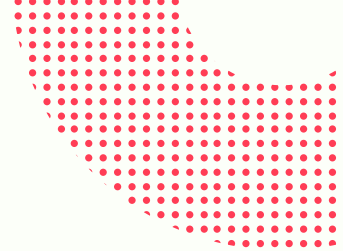


Above: Reactome ORA



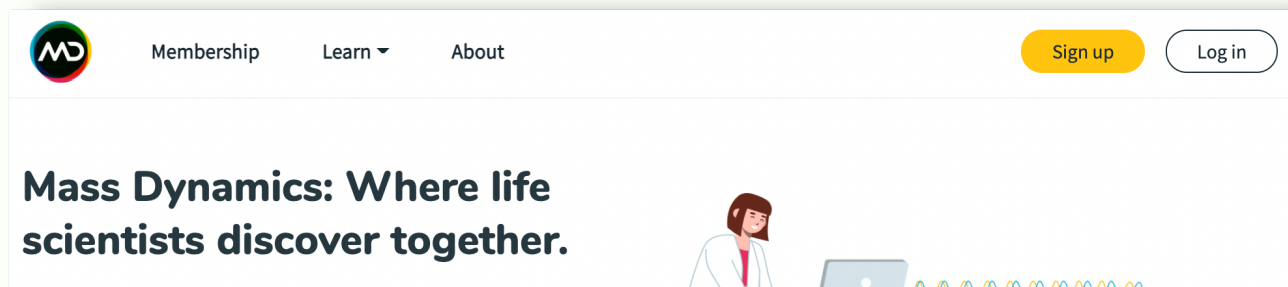
Above: Enrichment using Reactome or Gene Ontology (GO)

Above: Record insights



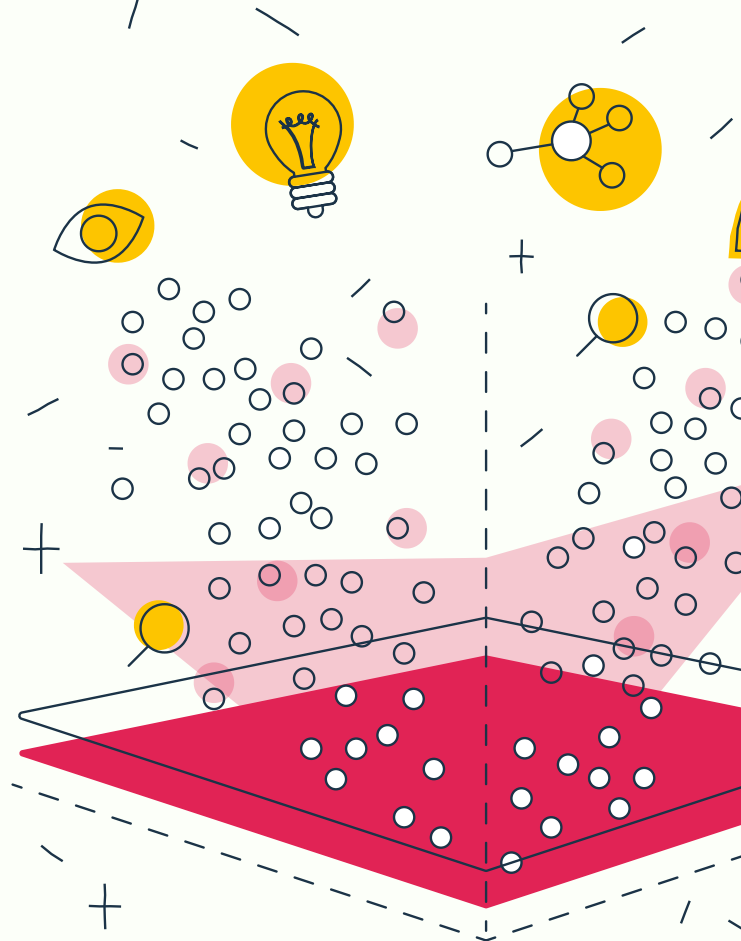
What now?

[Sign up](#) and try it out for yourself. You will gain access to example templates and datasets to explore.



Unsure if your data can be converted to the generic format?

With many tools being available to pre-process MSP data, it can be hard to fully anticipate how these data sources can be analysed once uploaded through generic format. We encourage you to get in touch with the Mass Dynamics team [here](#) if you have specific questions around your workflow.



About Mass Dynamics™

Our mission is to free humanity and society from the burden of disease by unlocking the magic of Mass Spectrometry (MS) and the power of existing biological knowledge. We do this by delivering a powerful software platform that seamlessly connects multi-disciplinary life scientists to answer biological questions and understand the building blocks of life - better, faster and easier.



www.massdynamics.com



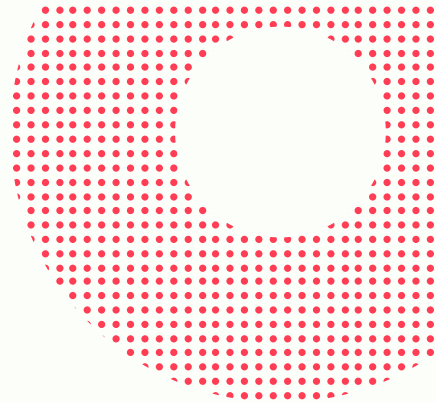
hello@massdynamics.com



github.com/massdynamics



Sign up for free membership, start today: app.massdynamics.com



References

1. Bloom, J., Triantafyllidis, A., Burton (Ngov), P., et. al. 2021. Mass Dynamics 1.0: A streamlined, web-based environment for analyzing, sharing and integrating Label-Free Data. bioRxiv 2021.03.03.433806; doi: <https://doi.org/10.1101/2021.03.03.433806>
2. Ritchie, ME., Phipson, B., Wu, D., Hu, Y., et. al. 2015. limma powers differential expression analyses for RNA-sequencing and microarray studies, *Nucleic Acids Research*, Volume 43, Issue 7, Page e47, <https://doi.org/10.1093/nar/gkv007>
3. Fabregat, A., Sidiropoulos, K., Viteri, G., Forner, O., et al. 2007. Reactome pathway analysis: a high-performance in-memory approach. *BMC Bioinformatics* 18, 142. <https://doi.org/10.1186/s12859-017-1559-2>
4. The Gene Ontology Consortium. 2019. "The Gene Ontology Resource: 20 Years and Still GOing Strong." *Nucleic Acids Research* 47 (D1): D330–38.