MultiAlign Tutorial 02 – Data Processing

BRIAN LAMARCHE

About this tutorial

- This tutorial will describe the data that MultiAlign processes.
- You will learn:
  - Terminology
  - How data is linked in the resulting relational database

Data Sources

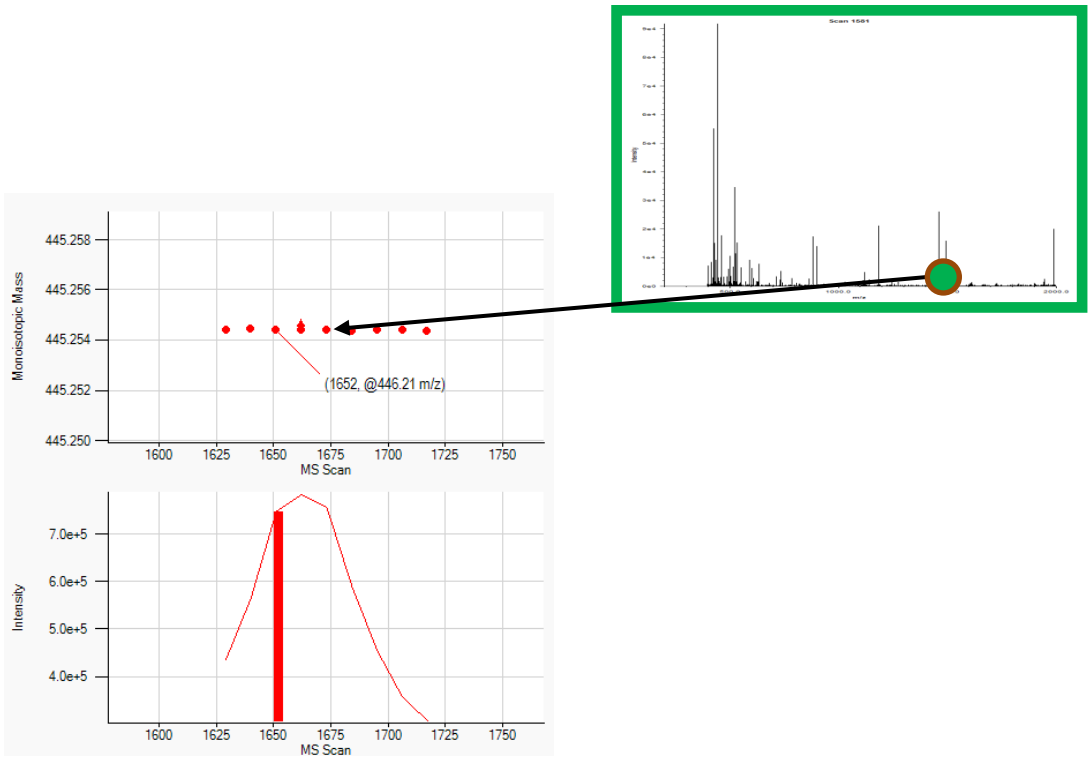
THIS SECTION PROVIDES THE BASICS OF HOW DATA IS DERIVED DURING THE MULTIALIGN PROCESS

Terminology – MS Feature

- MS Feature
  - A feature (isotopic distribution) that was identified in a single MS^1 scan from a single dataset

Mono Mass vs. Scan

LC-Elution Profile (monoisotopic peak)

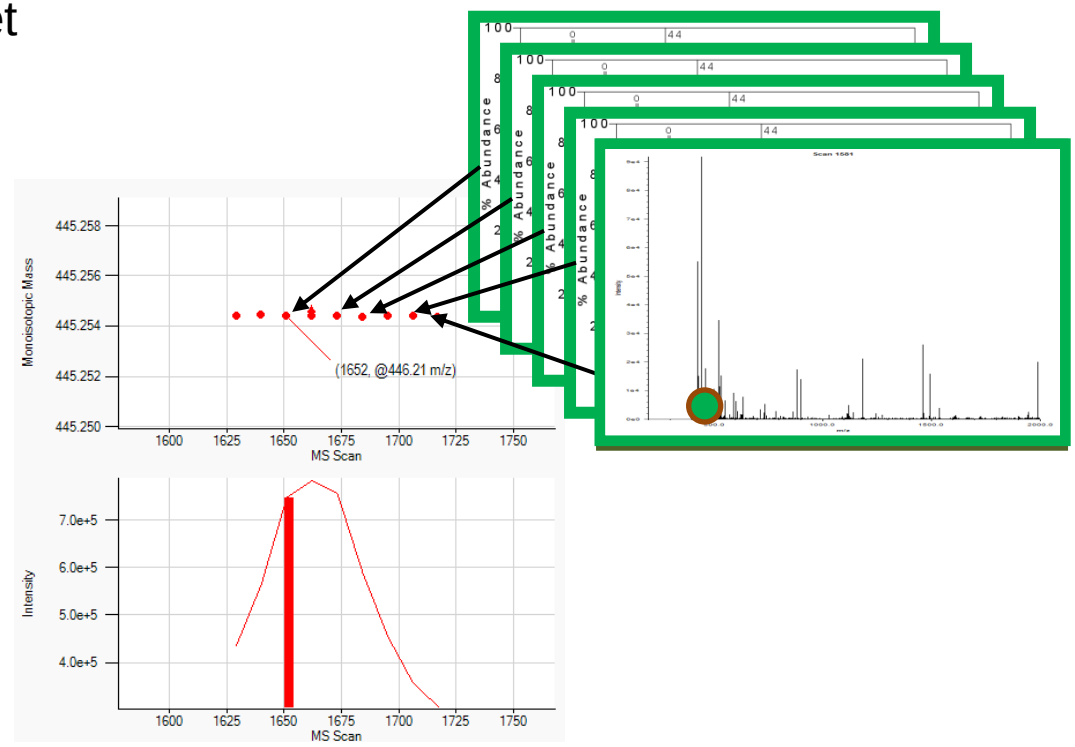


Terminology – LC-MS Feature

- MS Feature
  - A feature (isotopic distribution) that was identified in a single MS^1 scan from a single dataset
- LC-MS Feature
  - A collection of MS features that was grouped across multiple scans from a single dataset

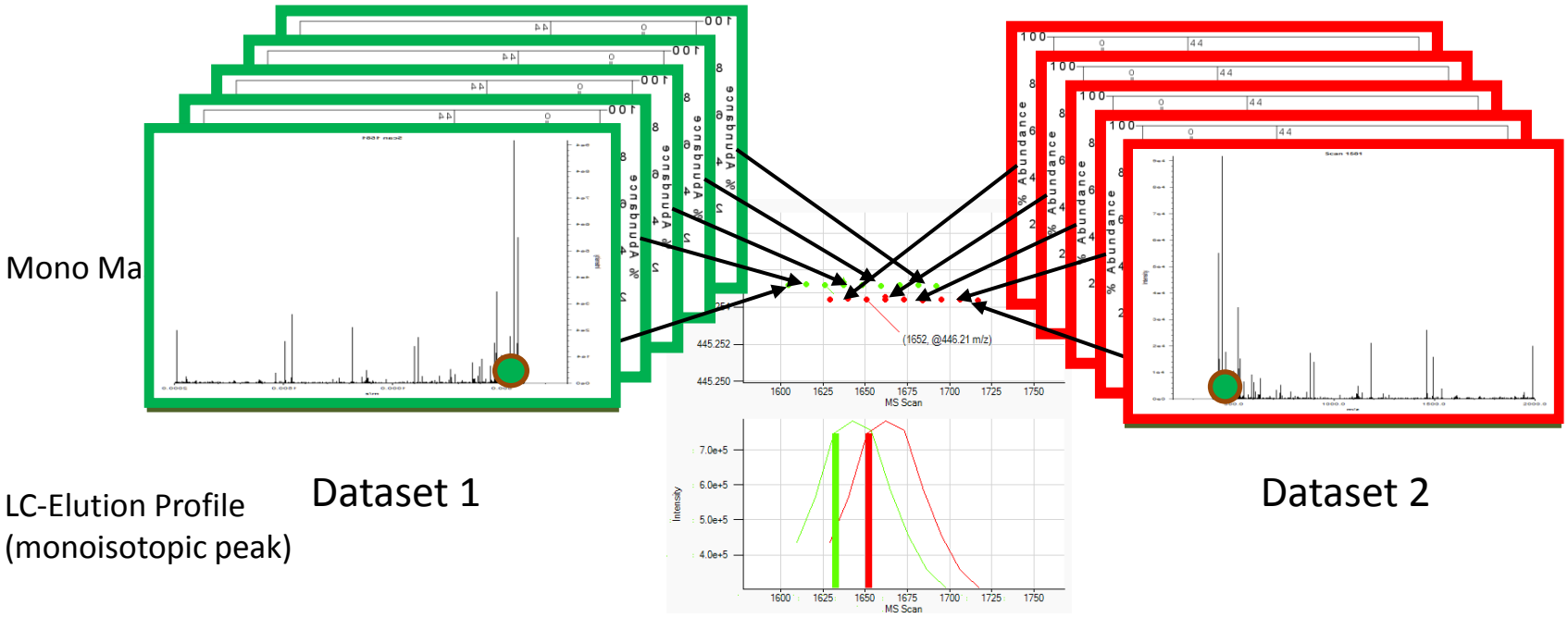
Mono Mass vs. Scan

LC-Elution Profile (monoisotopic peak)



Terminology – LC-MS Feature Cluster

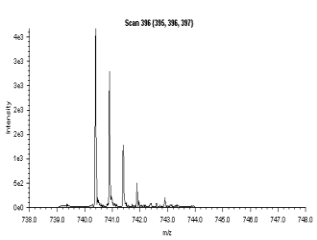
- MS Feature
  - A feature (isotopic distribution) that was identified in a single MS^1 scan
- LC-MS Feature
  - A collection of MS features that was grouped across multiple scans
- LC-MS Feature Cluster (consensus feature)
  - A collection of LC-MS Features that were found across datasets



Data Processing Workflow

THIS SECTION DESCRIBES HOW MULTIALIGN ANALYZES DATA IN COMPARISON TO VIPER

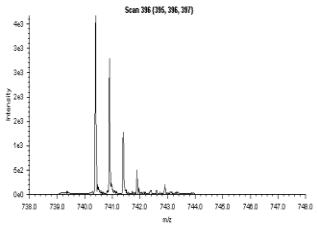
AMT tag approach – single dataset



Traditional workflows would be to:

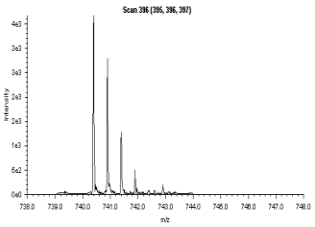
1. Deisotope an LC-MS instrument run into MS Features
2. Detect LC-MS Features
3. Align these features to a reference, i.e. mass tag database
4. Identify Peptides, i.e. Peak Match
5. Perform Quantitation using downstream tools

AMT tag approach variants



But sometimes AMT tag databases are not available because genomic information is unknown, as with some microbial communities or if performing metabolomic based analysis.

AMT tag approach variants



deisotoping (Decon2ls)

LC-MS Feature Finding

deisotoping (Decon2ls)

LC-MS Feature Finding

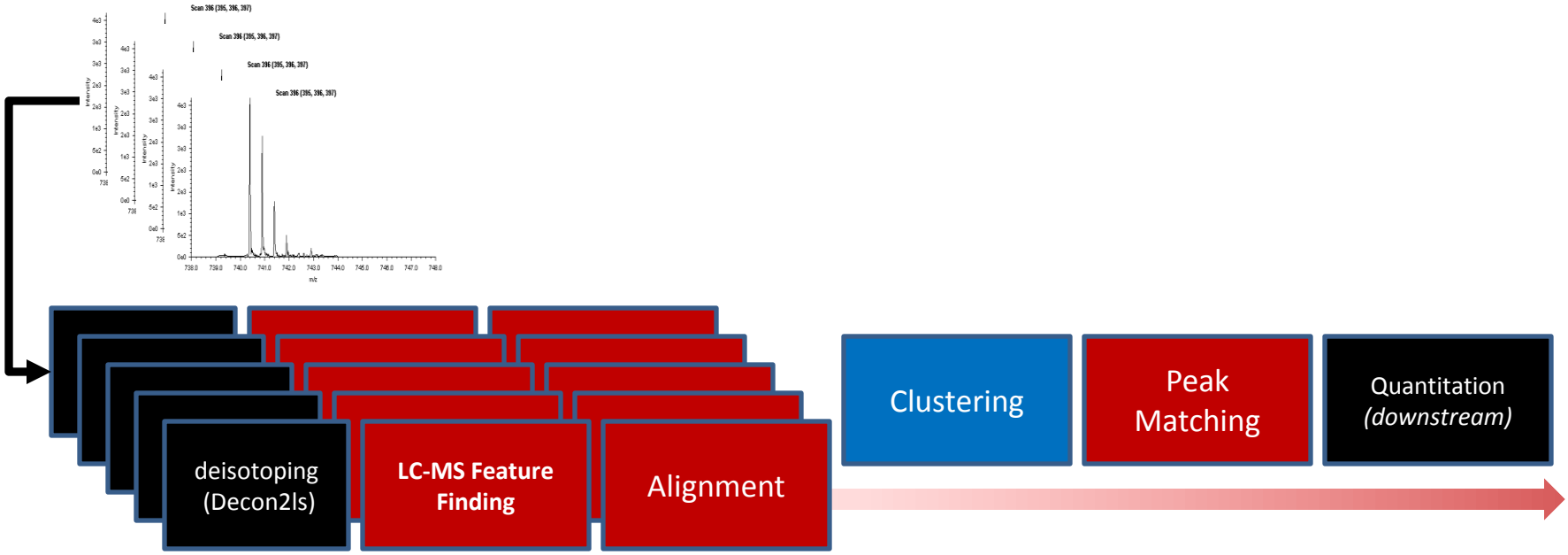
Alignment

Clustering

Quantitation *(downstream)*

Instead one may be interested in purely comparative dataset analysis. In this case, a dataset can be used as the reference for alignment.

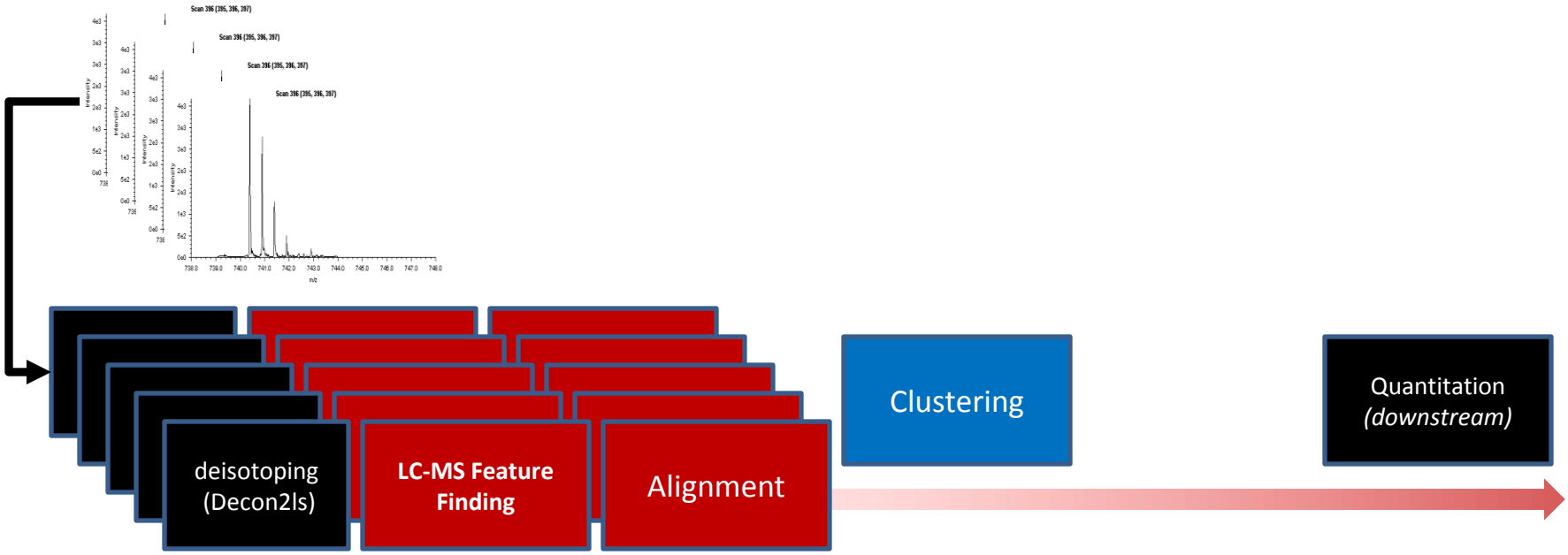
MultiAlign – Tool for Multiple dataset analysis



MultiAlign is a multiple dataset analysis tool (LC-MS, LC-IMS-MS) whose intention is to find common features across multiple datasets. It also can perform feature identification, e.g. for peptides) via the STAC algorithm.

MultiAlign includes an additional processing step clustering LC-MS features across datasets after alignment

MultiAlign – Tool for Multiple dataset analysis



MultiAlign can skip the identification (peak matching) step if no AMT tag database is available.

Traceback

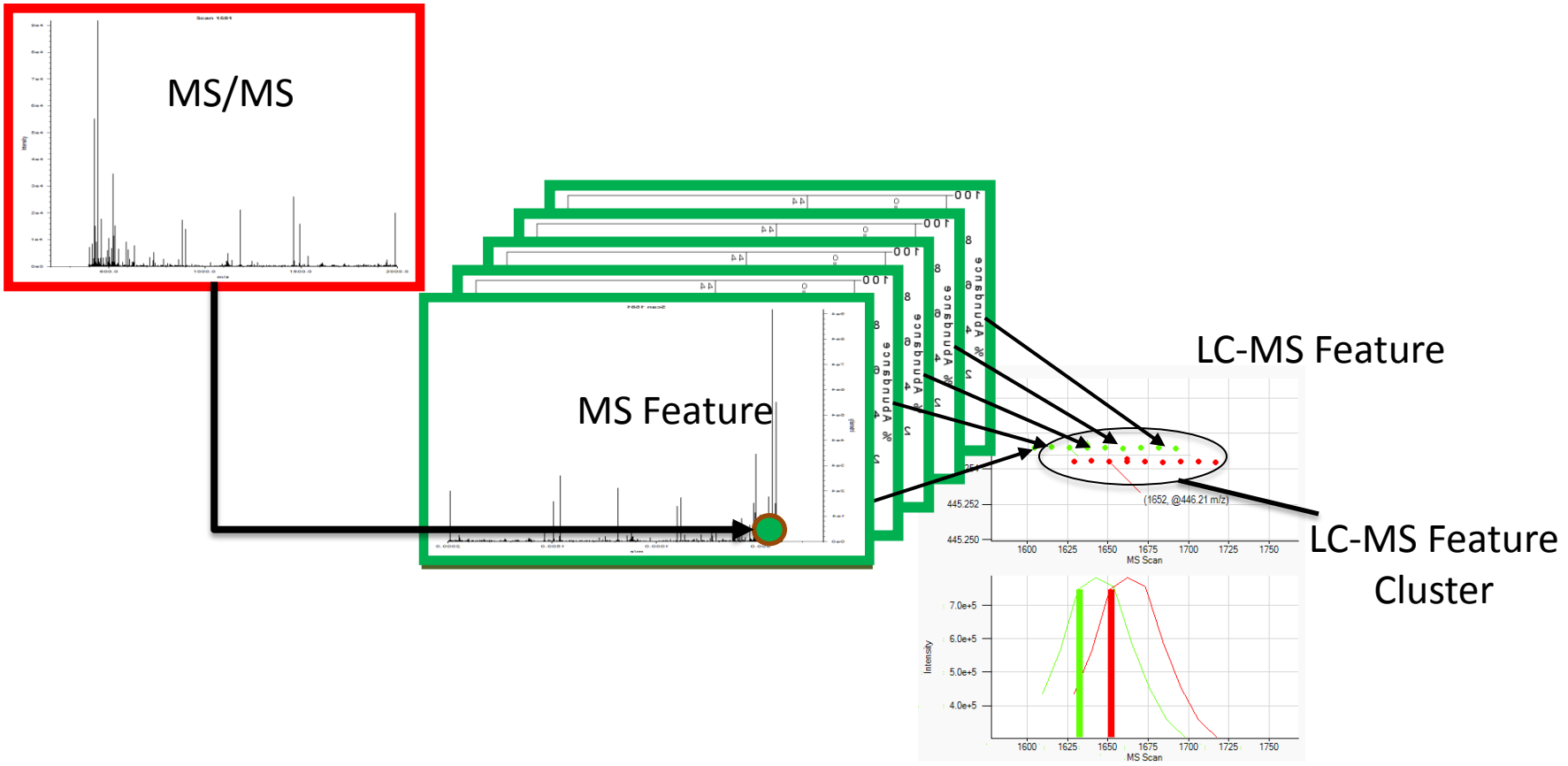
THIS SECTION DESCRIBES TRACEBACK AND HOW DATA IS LINKED TOGETHER IN THE MULTIALIGN RESULTS DATABASE

What is traceback?

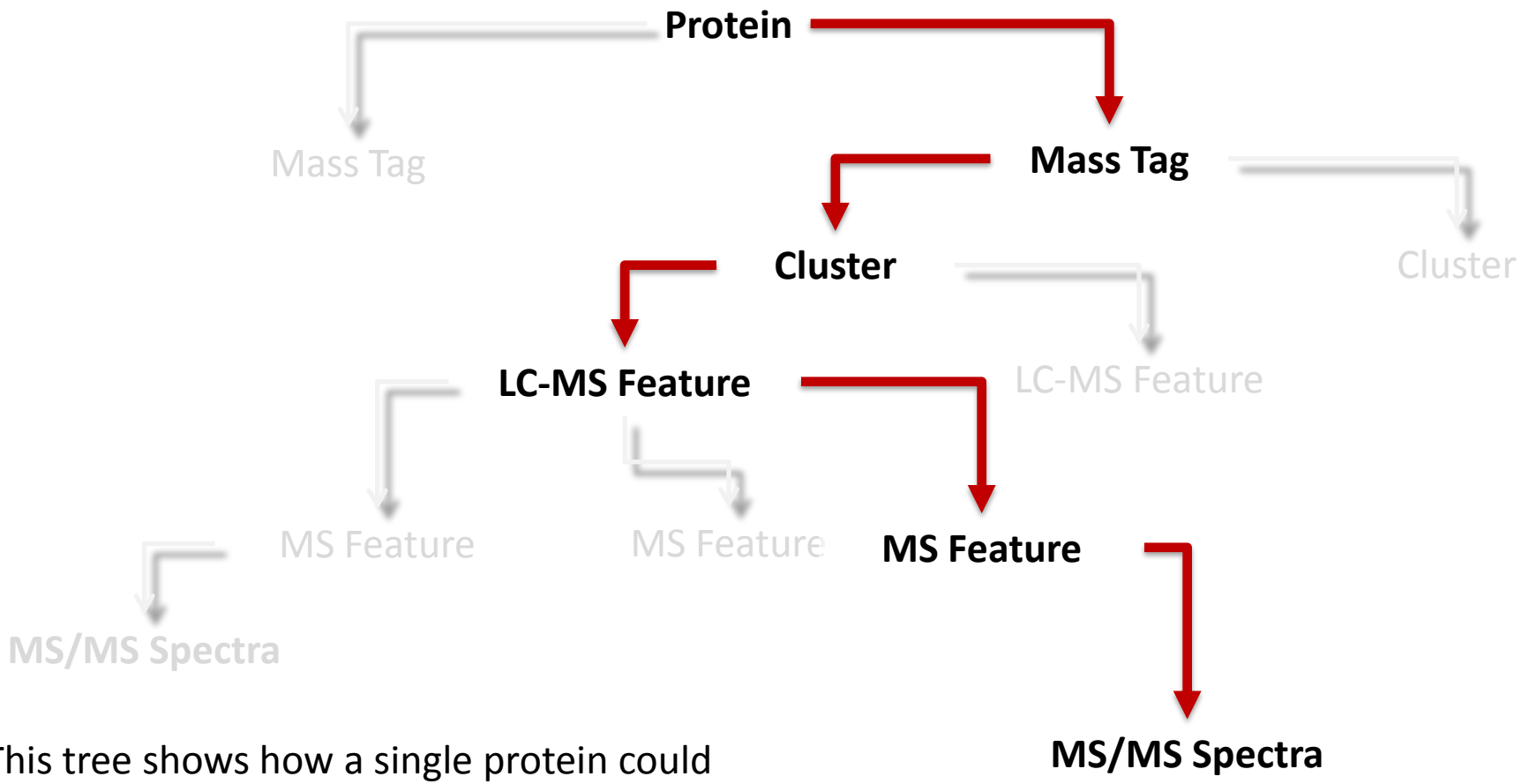
- Traceback simply implies the ability to start with a feature identification (e.g. peptide identification) or a cluster of features and trace all of the way to the raw spectra (MS^1) and MS/MS spectra if available.

Traceback Visually with spectra

With traceback we are trying to link a cluster, mass tag, or protein to their respective MS/MS spectra.



Mapping a protein to MS/MS spectra

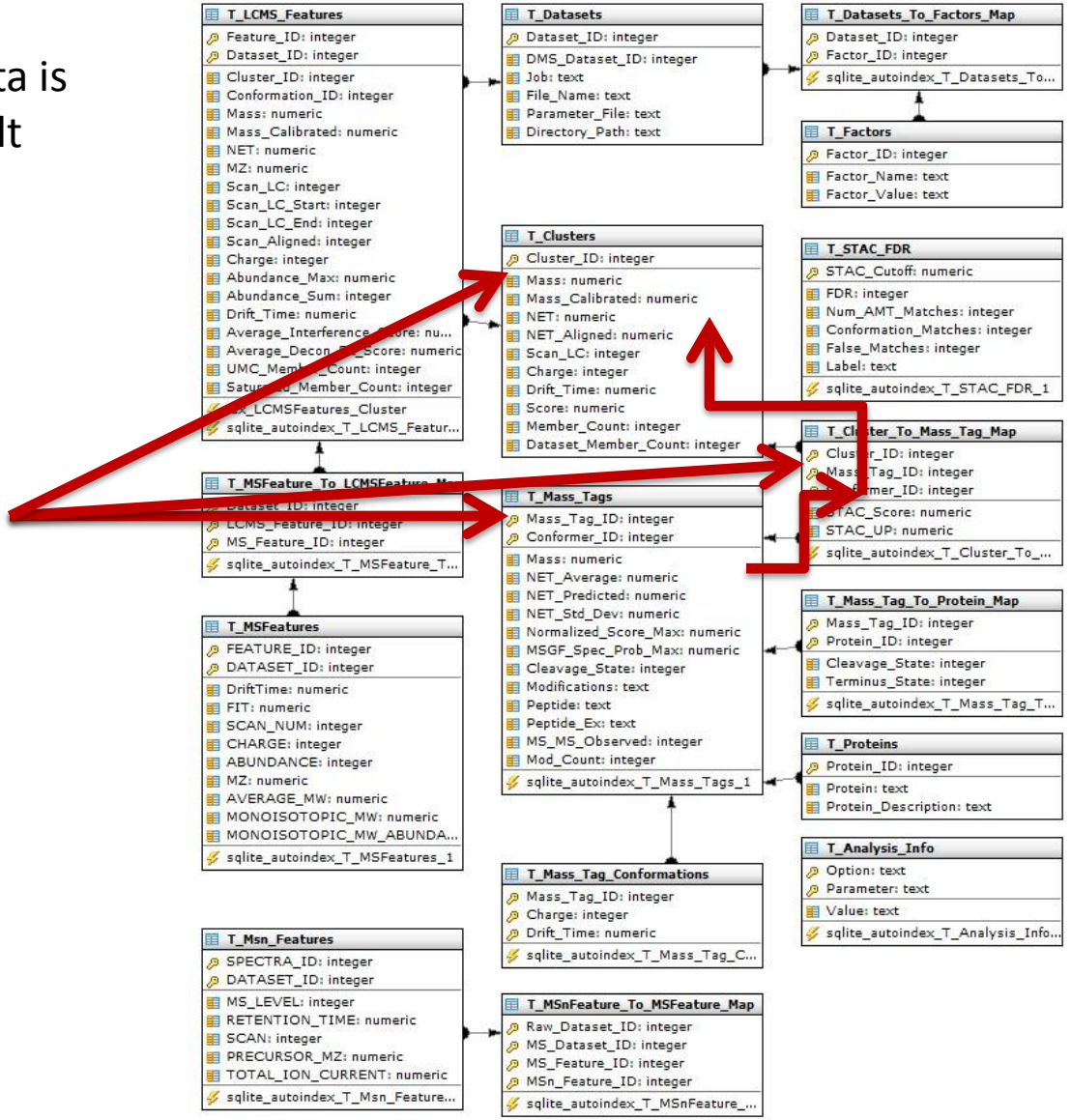


This tree shows how a single protein could map to an MS/MS spectra in one of the datasets analyzed.

Database Schema – Mapping from a mass tag to an MS/MS spectra

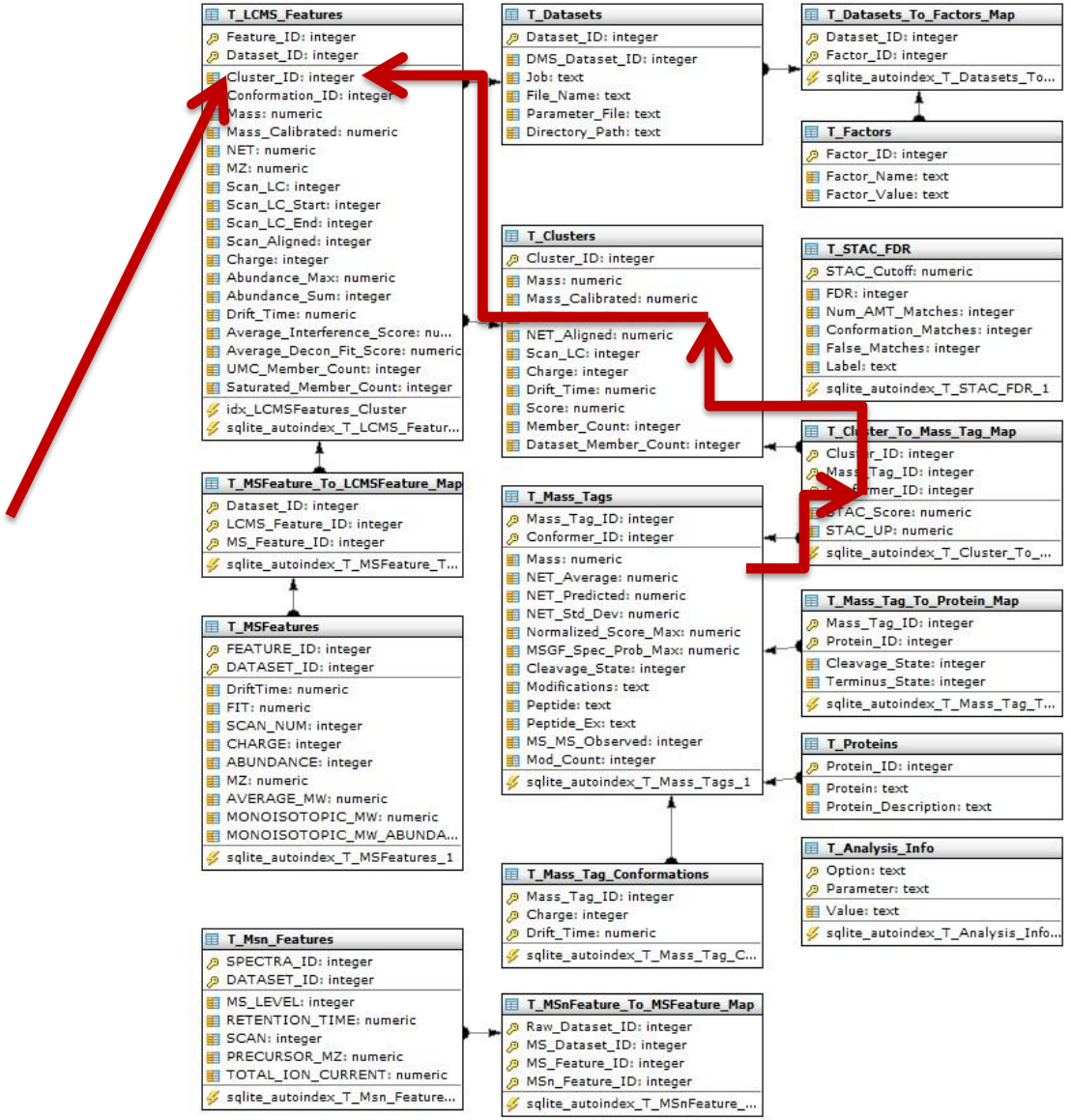
Here we show how this data is linked in the SQLite result database.

Starting with a mass tag, query data joining on mass tag id and cluster id through the mapping table



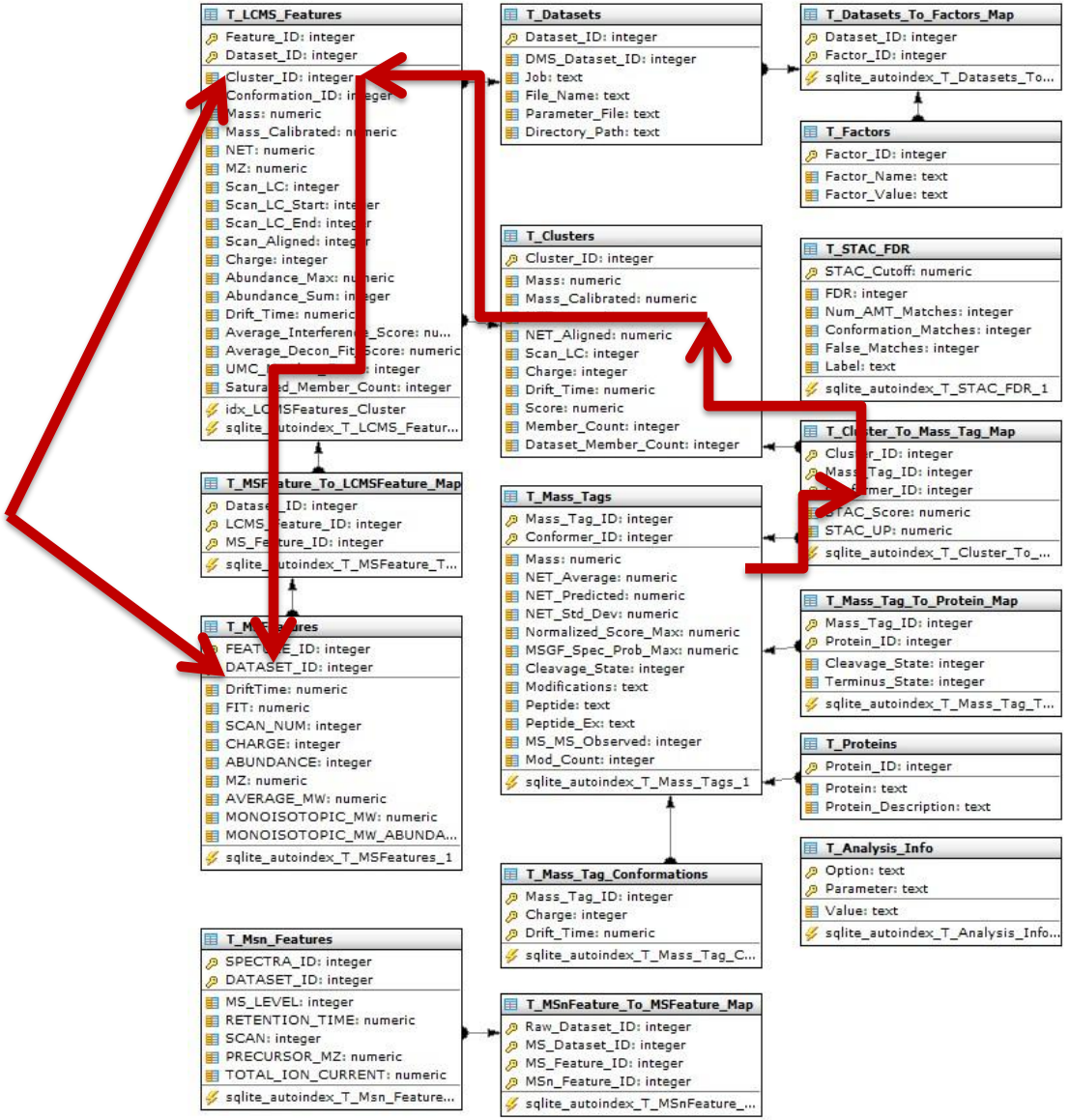
Database Schema – Mapping from a mass tag to an MS/MS spectra

Then map the cluster to all possible LC-MS features based on cluster id in the T_LCMS_Features Table



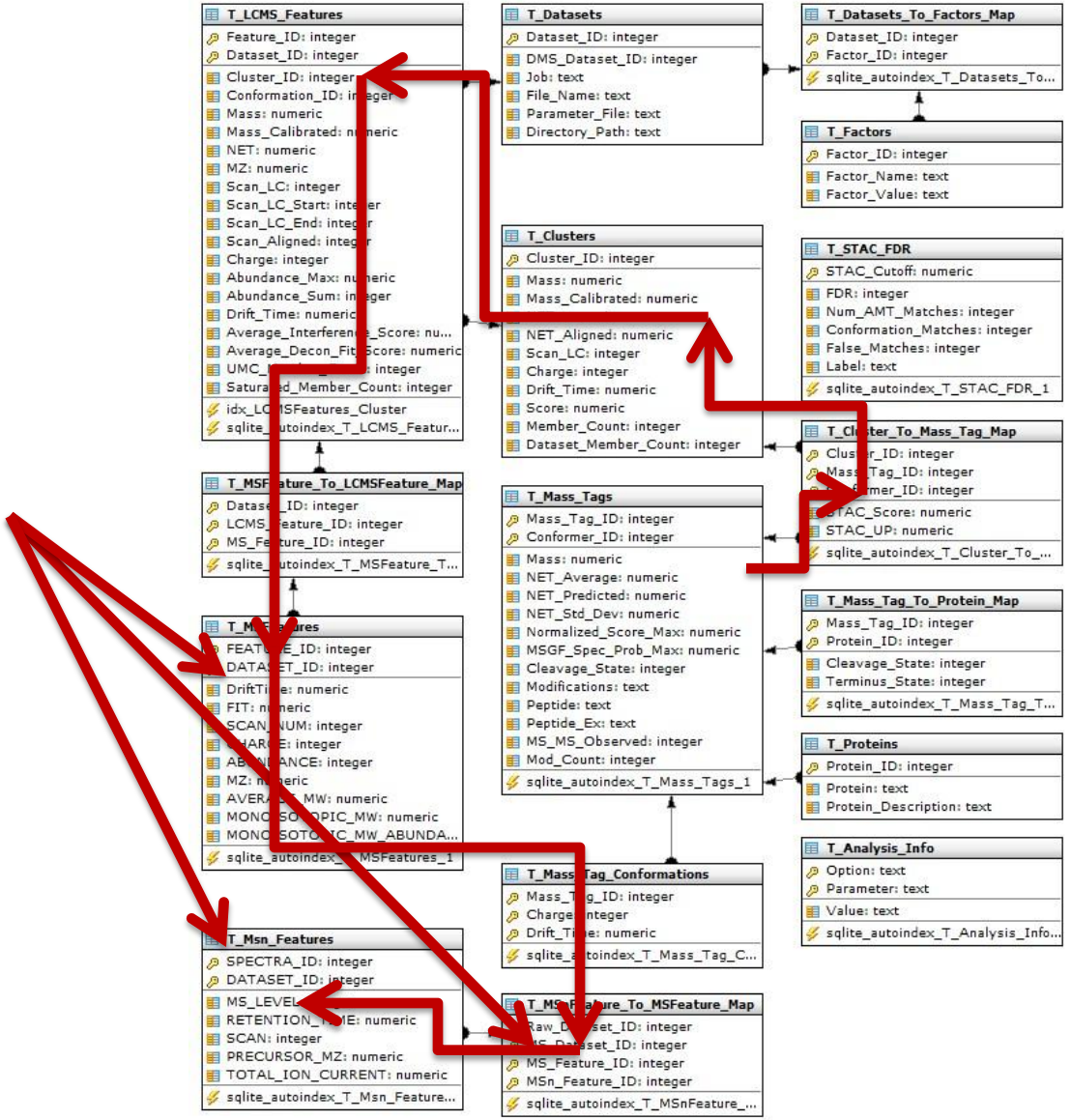
Database Schema – Mapping from a mass tag to an MS/MS spectra

Then map each LC-MS features to its appropriate set of MS Features based on Dataset ID and LC-MS feature ID



Database Schema – Mapping from a mass tag to an MS/MS spectra

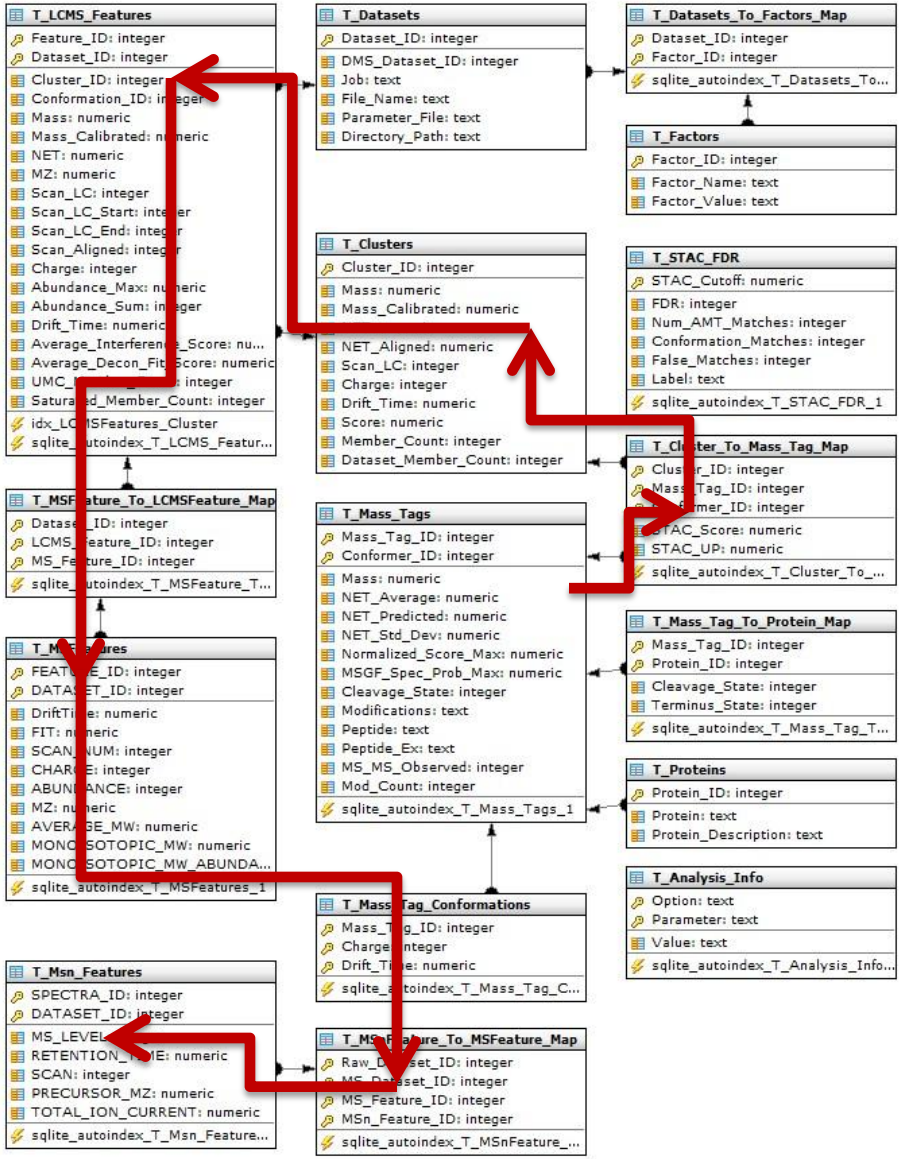
Then link every MS feature to an appropriate MSn Spectra ID through the T_MsnFeature_To_MSFeature_Map



Database Schema – Mapping from a mass tag to an MS/MS spectra

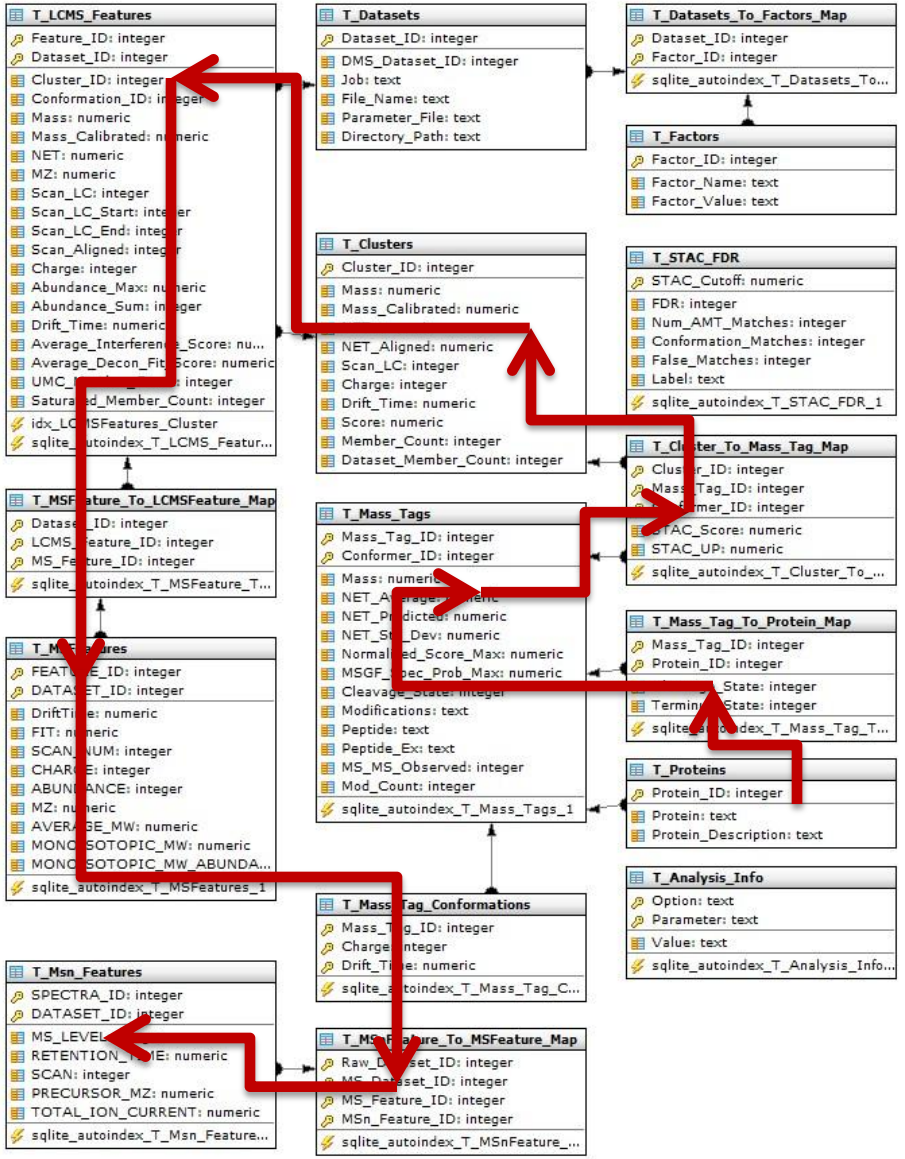
Now every identification is mapped back through to a possible MS/MS spectra.

If AMT is not performed, you can start from the T_Clusters step and trace backwards

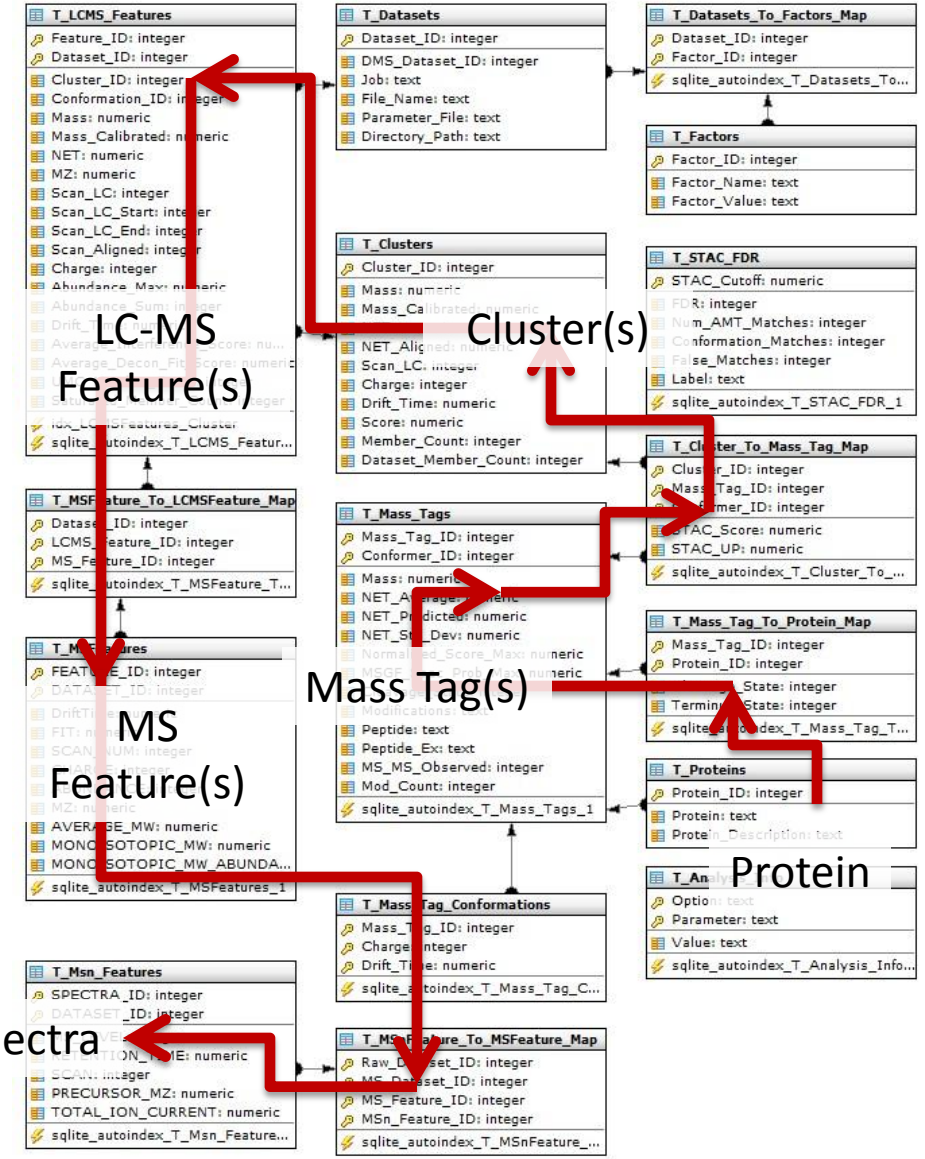


Database Schema – Mapping a protein to MS/MS spectra

Or you could start from a significantly changing protein ID and move backwards, this additional step is shown with the additional link through the T_Mass_Tag To_Protein_Map table



Database Schema – Mapping a protein to MS/MS spectra



- For more information see the MultiAlign website:

http://omics.pnl.gov/software/MultiAlign.php