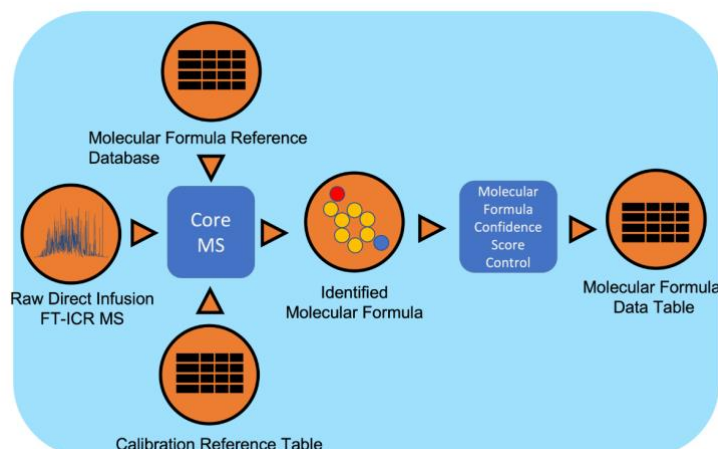


## Natural Organic Matter Workflow (v4.1.5)



### Overview

This workflow takes FTICR mass spectrometry data collected from organic extracts to determine the molecular formulas of natural organic biomolecules in the input sample.

### Running the Workflow

Currently, this workflow can be run in [NMDC EDGE](#) or from the command line (CLI instructions and requirements are found [here](#)).

### Input

The input for this workflow is the output from a massSpec experiment (a massSpec list) which includes a minimum of two columns of data: a mass-to-charge ratio ( $m/z$ ) and a signal intensity (Intensity) column for every feature in the analysis. A calibration file of molecular formula references is also required when running the workflow via command line (This calibration file is built into NMDC EDGE).

- **Acceptable file formats:** .raw, .tsv, .csv, .xlsx

### Details

Direct Infusion Fourier Transform Ion Cyclotron Resonance mass spectrometry (DI FTICR-MS) data undergoes signal processing and molecular formula assignment leveraging EMSL's CoreMS framework. Raw time domain data is transformed into the  $m/z$  domain using Fourier Transform and Ledford equation. Data is denoised followed by peak picking, recalibration using an external reference list of known compounds, and searched against a dynamically generated molecular formula library with a defined molecular search space. The confidence scores for all the molecular formula candidates are calculated based on the mass accuracy and fine isotopic structure, and the best candidate assigned as the highest score. This workflow will not work as reliably with Orbitrap mass spectrometry data.

### Software Versions

- CoreMS (2-clause BSD)
- Click (BSD 3-Clause "New" or "Revised" License)

### Output

The primary output file is the Molecular Formula Data Table (in a .csv file).

Primary Output Files	Description
INPUT_NAME.csv	m/z, Peak Height, Peak Area, Molecular Formula IDs, Confidence Score, etc.

## Running the Natural Organic Matter Workflow in NMDC EDGE

### Select a workflow

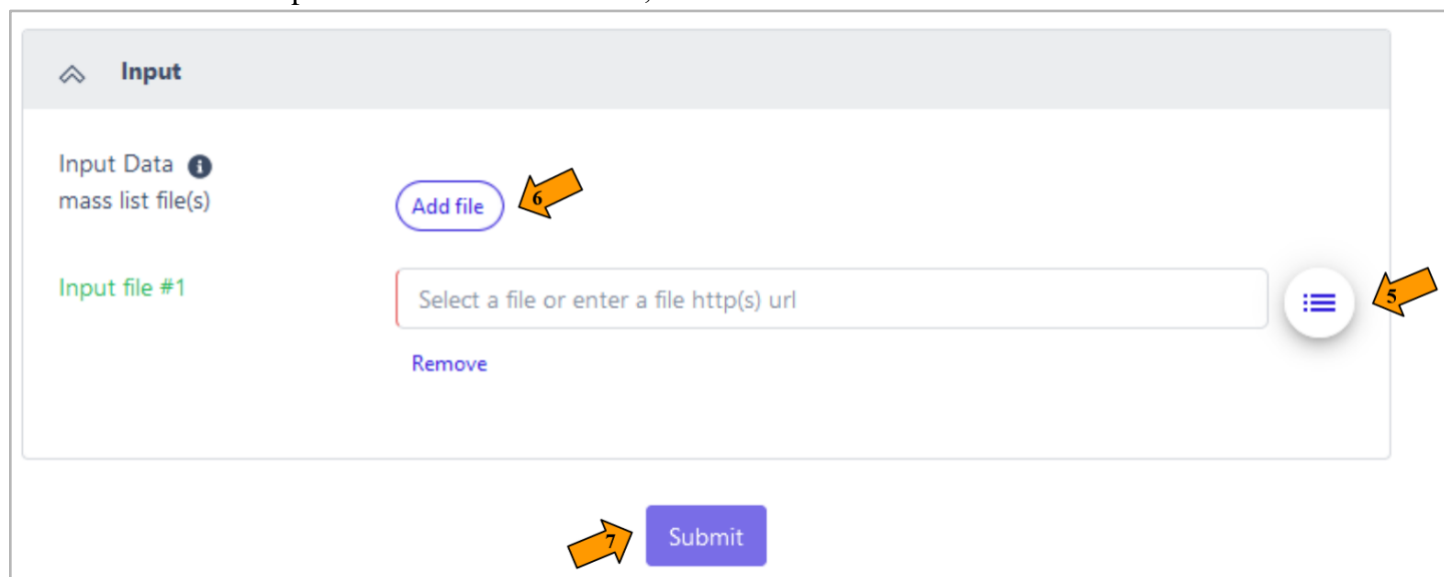
1. From the Organic Matter category in the left menu bar, select 'Run a Single Workflow'.
2. Enter a **unique** project name with no spaces (underscores are fine).
3. A description is optional, but helpful.
4. Select 'EnviroMS' from the dropdown menu under Workflow.

The screenshot shows the NMDC EDGE web interface. On the left is a dark sidebar menu with categories: Home, Tutorials, Public Projects, Upload Files, NMDC (containing Sample Submission Portal and Data Portal), and WORKFLOWS. Under WORKFLOWS, 'Organic Matter' is expanded, and 'Run a Single Workflow' is highlighted with an orange arrow labeled '1'. The main content area is titled 'Organic Matter | Run Single Workflow' and 'Run a Single Workflow'. It contains three input fields: 'Project/Run Name' (required, 3-30 characters) with an orange arrow labeled '2', 'Description' (optional) with an orange arrow labeled '3', and a 'Workflow' dropdown menu with 'EnviroMS' selected, indicated by an orange arrow labeled '4'. A 'Submit' button is at the bottom right.

## Input

The Natural Organic Matter workflow input is the output from a massSpec experiment (a massSpec list) with a minimum of two columns of data: a mass-to-charge ratio (m/z) and a signal intensity (Intensity) column for every feature in the analysis. **Acceptable file formats:** .tsv, .csv, .raw, .xlsx

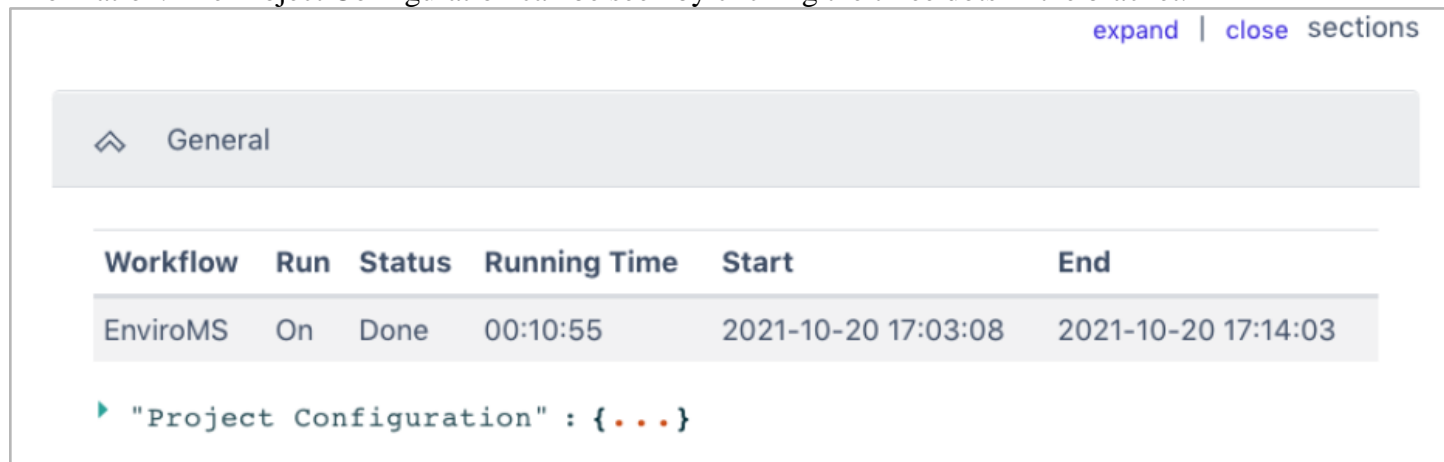
5. Click the button to the right of the input blank for data to select the data file for the analysis. (If there are separate files, there will be two input blanks.) A box called 'Select a File' will open to allow the user to find the desired file(s) from the public data folder or files uploaded by the user.
6. Additional input files can be added by clicking the 'Add file' button to create additional input blanks.
7. Once all the input files have been selected, click 'Submit'.



The screenshot shows the 'Input' section of a web interface. At the top, there is a header 'Input' with a chevron icon. Below it, there is a section for 'Input Data' with a sub-label 'mass list file(s)'. To the right of this is a blue 'Add file' button, which is pointed to by an orange arrow labeled '6'. Below the 'Add file' button is a text input field labeled 'Input file #1' with the placeholder text 'Select a file or enter a file http(s) url'. To the right of the input field is a circular menu icon, which is pointed to by an orange arrow labeled '5'. Below the input field is a blue 'Remove' button. At the bottom of the section is a large blue 'Submit' button, which is pointed to by an orange arrow labeled '7'.

## Output

The General section of the output shows which workflow and which tools were run and the run time information. The Project Configuration can be seen by clicking the three dots in the bracket.



The screenshot shows the 'General' section of the output. At the top right, there are links for 'expand' and 'close sections'. Below this is a header 'General' with a chevron icon. Below the header is a table with the following columns: Workflow, Run, Status, Running Time, Start, and End. The table contains one row for 'EnviroMS'. Below the table is a section for 'Project Configuration' with a bracket icon.

Workflow	Run	Status	Running Time	Start	End
EnviroMS	On	Done	00:10:55	2021-10-20 17:03:08	2021-10-20 17:14:03

► "Project Configuration" : {...}

The Browser/Download Output section provides output files available to download. The primary output files are: the Molecular Formula Data-Table (.csv file) containing m/z measurements, Peak height, Peak Area, Molecular Formula Identification, Ion Type, and Confidence Score.

File	Size	Last Modified
EnviroMS		
20190709_WK_CADY_Auto_S16_H1_Post_O5_1_01_36		
dbe_vs_c		
ms_class		
mz_error_class		
van_krevelen		
20190709_WK_CADY_Auto_S16_H1_Post_O5_1_01_36.csv	4.93 MB	6 days ago
20190709_WK_CADY_Auto_S16_H1_Post_O5_1_01_36.json	9 kB	6 days ago
assigned_unassigned.png	13 kB	6 days ago
mz_error.png	73 kB	6 days ago

