DecoID User Interface Help Page

Updated 10/26/20

Questions: estancliffe@wustl.edu

Overview:

DecoID offers deconvolution and searching of MS/MS spectra. In its simplest form, it averages all spectra acquired for a particular feature of interest and then searches those spectra against a metabolomic database (mzCloud, MoNA, HMDB) based on accurate mass and similarity to the MS/MS spectrum in the database. In addition, "chimeric" MS/MS spectra can be deconvolved, where, instead of searching against each database spectrum individually, all possible linear combinations of database spectra are considered and the combination that results in the most similar spectrum to the acquired spectrum is used to improve the identification. For details on how DecoID works, please consult the publication. In this help page, the parameters available in the user interface are detailed and examples for DDA and DIA data setup and analysis are given.

Installation:

A standalone GUI built for 64 bit Windows systems can be downloaded here. After download, extract the contents from the .tar.gz file. Navigate into the extracted folder and find DecoIDGUI.exe. If you are using a machine other than 64 bit Windows, you will need to do run from the source code. First clone or download the DecoID GitHub project, the release can be found here. Python 3.7 will need to be installed. A download can be found here. Next install the DecoID package with pip. For instructions, see the DecoID GitHub page. You will also need to download the MoNA database file packaged as an asset in the DecoID release found here. The MoNA-export-LC-MS-MS_Spectra.db file will need to be moved to the DecoID/databases/ directory.

You may also need to install the Python module matplotlib which can be installed with the command:

```
>pip install matplotlib
```

Next, in DecoID/GUI/ run the command:

```
>python decoIDGUI.py
```

Currently, PyInstaller only supports Python version 3.5-3.7. Further, building the standalone GUI on macOS is currently not supported due to known bugs within PyInstaller. However, the source code that includes both the GUI and the Python module can be used across all operating systems.

Search Parameters:

Below the search parameters are defined:

- **Processor Number**- # of concurrent processes to run at once. In other words, how many spectra do you want to search and deconvolve in parallel. This number should be set as high as you can afford based on the hardware. On most laptop 2-3 is a good choice. This will still leave you with enough computing power for other uses. On a normal desktop, 5-6 would likely work well. On a workstation computer, 10-15 will give fast performance. As a general rule, I try to set it to the # of processors in the computer – 1. However, when using large reference databases (like MoNA)

with DIA data or MS/MS acquired with wide isolation windows, memory (RAM) usage may be high. To reduce possible memory limitations, the number of processors can be set to be lower. This generally reduces memory usage linearly.

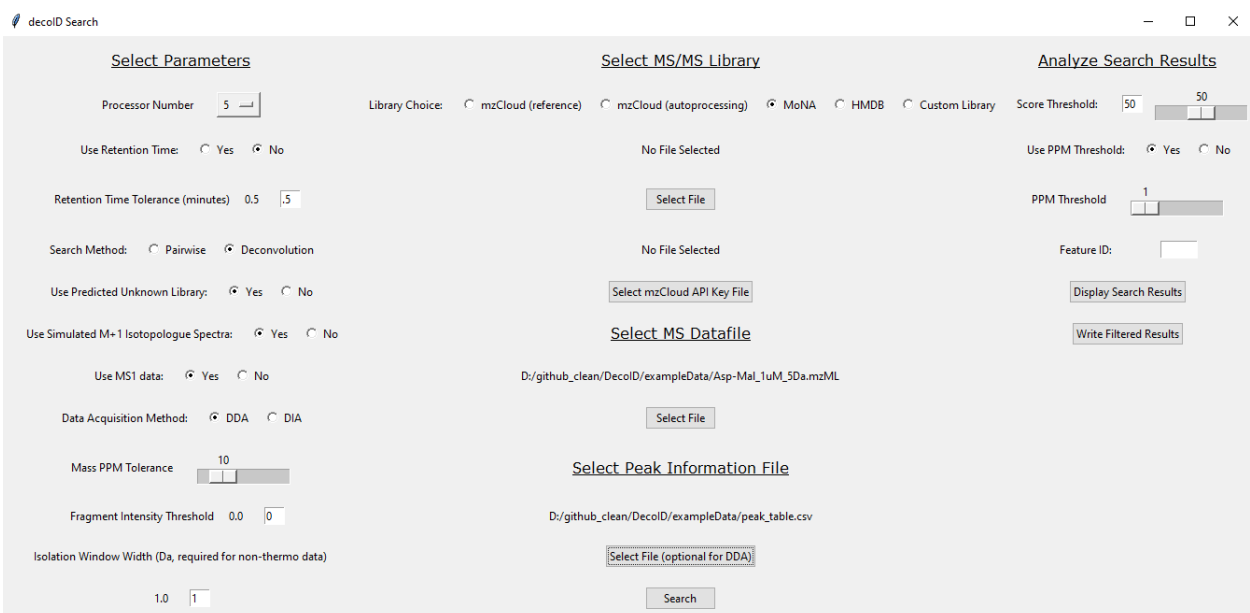
- **Use Retention Time**- If toggled "Yes", candidate spectra from the reference database will only be used to deconvolve the query spectra if the database retention time is either not known (-1 is database file) or the listed retention time is within the user defined tolerance. If "No", all retention times will be ignored.
- **Retention Time Tolerance**- Tolerance for retention time constrained deconvolution. Only meaningful if "Use Retention Time" is toggled to "Yes". This is the time in minutes where if the database retention differs from the retention time of the acquired query spectrum by more than this amount, it will not be used to deconvolve the acquired spectra
- **Search Method**- Pairwise: without deconvolution, simply performs spectral averaging and direct MS/MS search. Deconvolution: try to find an optimal linear combination of database spectra. If you suspect your data is chimeric (large isolation window, complex sample matrix), this will improve performance dramatically, but is slower than direct searching.
- **Use Predicted Unknown Library**- If toggled to "Yes", spectra with no match in the selected database, but are not chimeric, are used to deconvolve other spectra that were acquired. Usage of these spectra is restricted by the retention time of the "unknown" feature and the accurate mass. More details and examples can be found in the publication as the "on-the-fly unknown library". This can only be used in DDA workflows where MS1 full scan data is acquired along with the MS/MS spectra.
- **Use simulated M+1 isotopologue spectra**- If toggled "Yes", MS/MS fragments coming from orphan isotopologues are able to be deconvolved from a chimeric spectrum. This sort of contamination occurs when the parent M+0 is excluded from the isolation window, but the M+1 is not. The publication gives further details. In short, the M+1 isotopologue's MS/MS spectrum is modeled as a mixture of isotopomer spectra. Probable spectra for these isotopomers are predicted by shifting the top 5 most intense fragments in the M+0 spectrum by +1.003 m/z. This can only be used if MS1 full scan data is acquired along with the MS/MS, as is common in DDA and DIA workflows.
- **Use MS1 data**- If toggled "Yes" and MS1 data was collected in addition to MS/MS, the nearest MS1 spectrum to each acquired MS/MS spectrum is used to filter which compounds to use with deconvolution by requiring that each compound used to reconstruct the chimeric MS/MS spectrum have a peak at the corresponding precursor's m/z. This parameter is meaningless if "pairwise" is selected.
- **Data acquisition method**- Data dependent acquisition (DDA) for datasets where narrow isolation windows are used to target specific m/z values. In the DDA analysis workflow, the goal will be to identify the compounds targeted for MS/MS. In data independent acquisition (DIA). The goal is to identify all compounds in each spectrum which is assumed to be multiplexed. A common DIA workflow would be sequential acquisition of the entire m/z axis with 20 m/z isolation windows. A peak list is required for DIA (see below). Select the method that most closely matches your desired output and acquired data.
- **Mass PPM Tolerance-** This parameter determines the MS1 mass accuracy tolerance. This is used to assign spectra to features as well as filter database matches. This should be set to the mass accuracy of the instrument.

- **Fragment Intensity Threshold**- Absolute intensity threshold for MS/MS peaks. Any fragment ion below this threshold will be removed from the query spectra. This can be useful for noisy spectra with high background.
- **Isolation Window Width-** This parameter only needs to be entered for non-Thermo data. The value should be the width of the isolation window around each DDA spectrum (i.e. if the isolation window is 112-115 m/z, the entry should be 3.0).

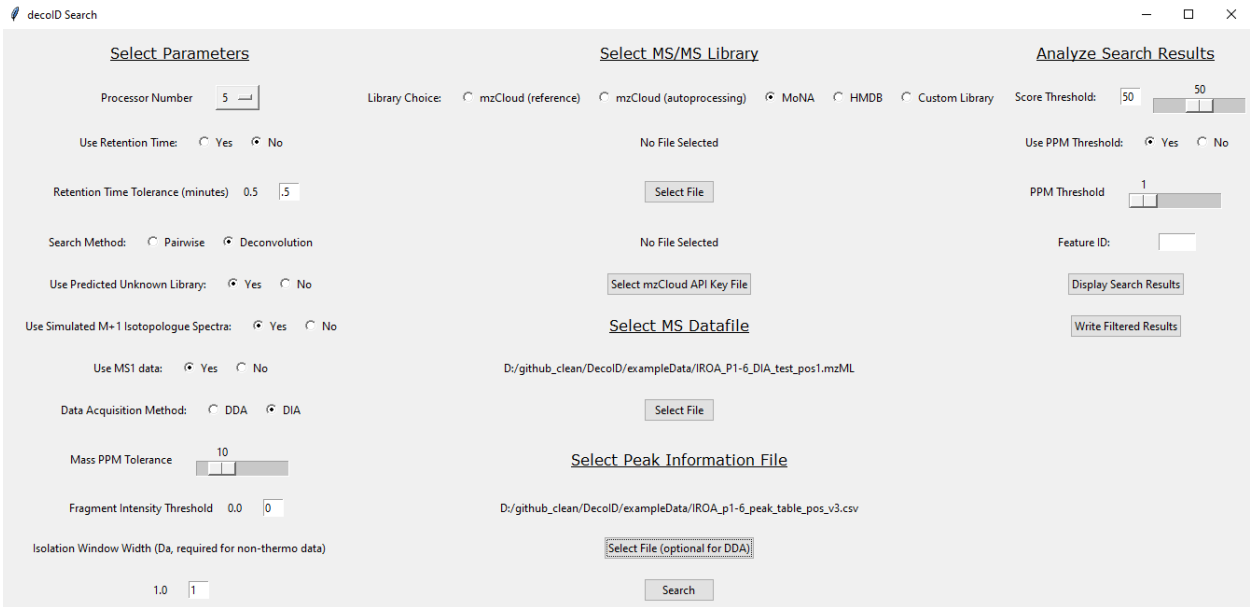
Data Input:

- **Select MS/MS Library-** This selects the database to use to identify and deconvolve the MS/MS spectra. MoNA, HMDB, and mzCloud are built into the software and can easily be selected. Usage of mzCloud requires input of an API key that can be obtained from Thermo-Fischer Scientific. Alternatively, a user defined database can be selected via the file selection dialog box. This file must be formatted as a NIST .msp file. For details on the formatting see the API documentation here.
- **Select mzCloud API Key File-** If usage of mzCloud library is desired, the text file containing the API key must be selected. For acquisition of the API key, contact Thermo-Fisher Scientific.
- **Select MS Datafile-** This is where the file containing the MS/MS information must be selected. If MS-Convert is installed and msconvert.exe is added to the system PATH, vendor formatted files can be selected and will be internally converted. Alternatively, DecoID accepts .mzML files that have already been converted and contain centroid data.
- **Select Peak Information-** This field is optional for DDA data and required for DIA data. Its purpose is to tell DecoID how to group and assign MS/MS spectra for each feature. This allows for spectral averaging and also aids downstream analyses. The file must contain the m/z and retention time bounds for each feature in a .csv file. Details on the format can be found here.

Example Setup for DDA Data:



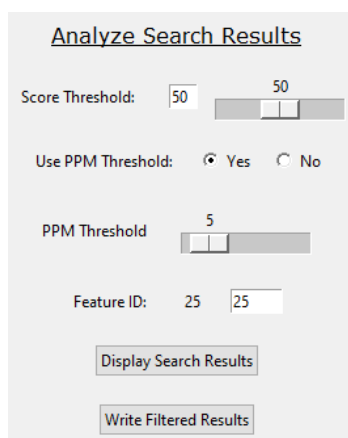
Example Setup for DIA Data:



Visualization Parameters:

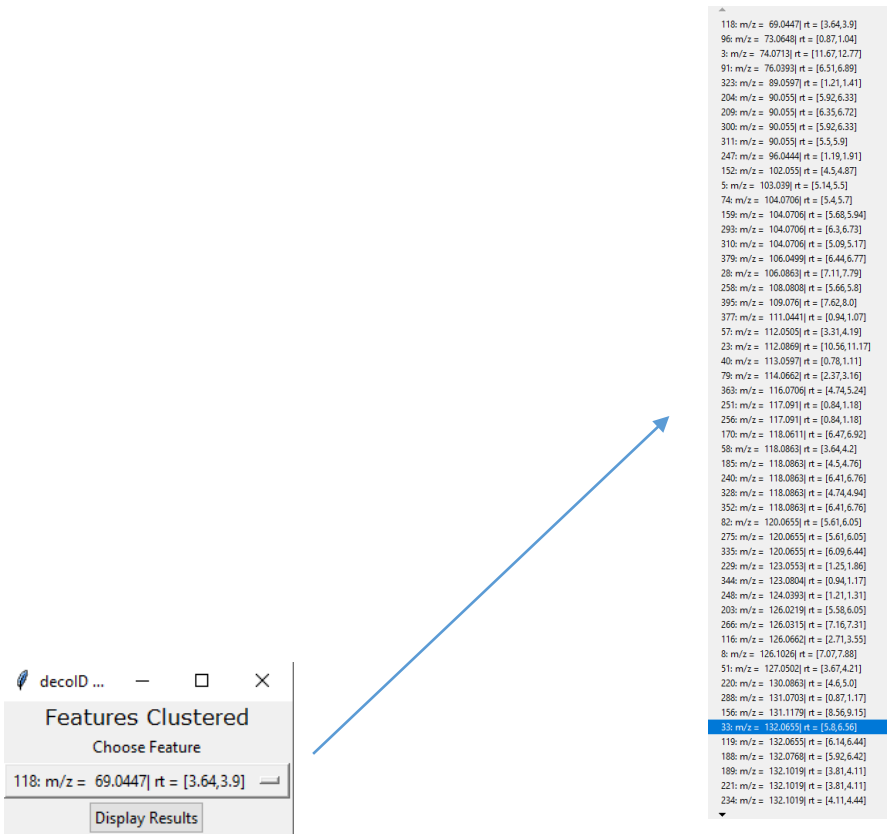
- **Score Threshold-** Minimum dot product similarity between database spectrum and output spectrum to be shown as a hit.
- **Use PPM Threshold-** Use a mass error tolerance threshold to filter results.
- **PPM Threshold-** Maximum allowed precursor mass error between database match and feature of interest m/z.
- **Feature ID-** If only interested in results for a particular feature of interest, enter the row number of the feature of interest from the peak information file.
- **Write Filtered Results-** Write a filtered list of results based on the thresholds given.

As an example the parameters to the right will give results for the feature in row 25 of the peak information file that have a dot product similarity of greater than 50 and with a mass error of less than 5 ppm.

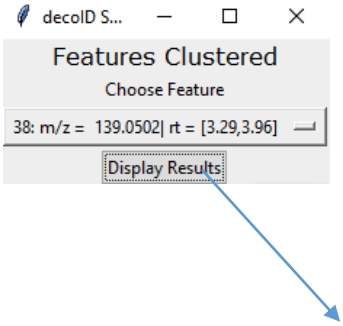


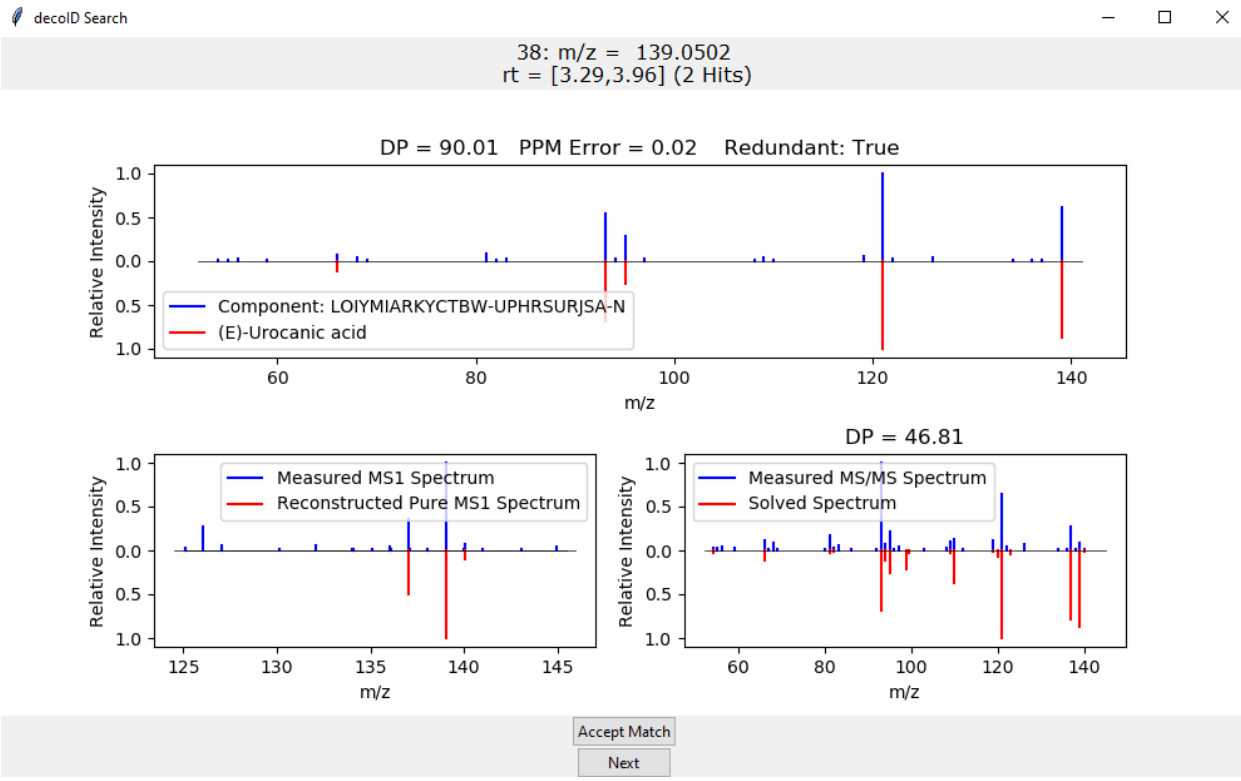
Visualization:

After clicking "Display Search Results" a drop down menu will be displayed that allows for a feature of interest to be selected. The precursor m/z and retention time bounds for each feature are listed.



After selecting a feature of interest, the results are display in order of match quality (highest->lowest).





In the visualization pane, the top panel gives the match with the deconvolved component of the acquired spectrum in blue and the database match in red. DP gives the dot product similarity and the PPM error gives the precursor mass error. If Redundant is listedas true. The matches component is not unique in that a different database compound could have been used to deconvolve the query spectrum. If redundant is false the component is unique for the deconvolution. The bottom left shows the MS1 spectrum around the isolation window used to acquire the MS/MS spectrum. The blue is the measured spectrum the red is what was reconstructed. The bottom right shows the entire measured MS/MS spectrum in blue and reconstructed spectrum in red that is found after deconvolution. The similarity between these spectra is given by the DP. For manual validation of matches, if a match is accepted, the "Accept Match" button can be pressed. This match will then be written to a separate output file given

with ending "_annotated.csv" and will follow the same format as the the "_decoID.csv" output file. Details on the output files produced by DecoID are given here.