The code included with this manuscript (“*In silico* MS/MS spectra for identifying unknowns: A critical examination using CFM-ID algorithms and ENTACT mixture samples”) is separated into two main groups, search code and results processing code. The documentation below is meant to explain the functionality of both sets of code as well as assist in both running and interpreting output of the code.

**Search Code:**

What it does:

1. Takes in mgf file as input (Instructions for creating MGF file from MS2 data below)
   1. Open MS2 file within Agilent MassHunter software
   2. File 🡪 Export 🡪 as MGF…
   3. Export Contents 🡪 Entire data file
   4. Click OK
2. Converts mgf file into a csv format (de-duplicates spectra by precursor mass, keeping spectrum with highest sum intensity of ions)
3. Searches CFM-ID database for candidate compounds based on mass window around precursor mass. Iterates through all precursor masses. Pulls back CFM-ID spectra (CE 10, 20, 40) associated with each candidate compound
4. Scores each candidate compound CFM-ID spectrum against experimental spectrum using dot-cosine product algorithm
5. Returns output csv file containing candidate compounds and their associated CFM-ID scores for all MS2 spectra

Scripts:

* Run\_v24.py – Main script (Run this)
  + Within this script, select the directory containing the experimental mgf files as well as the name of the mgf files to search.
* CosineDotProduct\_v24.py – Script containing sub-functions
* Mgf\_parser\_v24.py – Script containing sub-functions

Output:

* Excel file: <MGF filename>\_CFMID\_Multiscores\_AllHits.xlsx
* Column legend:
  + MASS – monoisotopic mass of candidate compound
  + DTXCID – DSSTox Chemical Identifier of candidate compound
  + FORMULA – molecular formula of candidate compound
  + Energy0 – Dot-product cosine score of experimental spectrum against candidate compound CFM-ID spectrum (predicted at CE 10)
  + Energy1 – Dot-product cosine score of experimental spectrum against candidate compound CFM-ID spectrum (predicted at CE 20)
  + Energy2 – Dot-product cosine score of experimental spectrum against candidate compound CFM-ID spectrum (predicted at CE 40)
  + RANK\_E0 – Rank of candidate compound (based on energy0 score) within all candidate compounds with the same molecular formula
  + RANK\_E1 – Rank of candidate compound (based on energy1 score) within all candidate compounds with the same molecular formula
  + RANK\_E2 – Rank of candidate compound (based on energy2 score) within all candidate compounds with the same molecular formula
  + MATCHES – Number of candidate compounds with that specific molecular formula
  + MASS\_in\_MGF – Experimental precursor mass from mgf file
* Excel file: <MGF filename>\_CFMID\_OneScore\_AllHits.xlsx
* This file is the same as above, except scores and ranks have been aggregated between the three CE predicted spectra
* Column legend below:
  + SCORE - Dot-product cosine scores of experimental spectrum against candidate compound CFM-ID spectrum (summed across all three predicted CE scores)
  + RANK - Rank of candidate compound (based on SCORE) within all candidate compounds with the same molecular formula

**Results Processing Code:**

What it does:

1. Reads in CFM-ID search results: excel files outputted from search code (see above)
2. Reads in previous ENTACT analysis results for each ENTACT mixture
3. Out of all ENTACT mixture compounds (#2), select only ENTACT mixture compounds which were previously confidently identified (‘Decision’ column = ‘Pass’)
4. Merges in CFM-ID search results (#1) with ENTACT mixture compounds previously identified (#3)
5. Within each set of CFM-ID search results for each ENTACT mixture compound, calculate quotient and percentile values for all candidate compounds
6. Returns output file showing CFM-ID scores, ranks, score quotients, score percentile values for all candidate compounds only corresponding to ENTACT mixture compounds

Scripts:

* ENTACT v1 CFMID analysis (Approach 1).py
* ENTACT v1 CFMID analysis (Approach 2).py
* ENTACT v1 CFMID analysis (Approach 3).py
* ENTACT v1 CFMID blinded analysis of data.py

Output:

* ENTACT v1 CFMID analysis (Approach 1).py
  + Excel files:
    - Approach 1 exp CE10 ENTACT compound cfmid results.csv
    - Approach 1 exp CE20 ENTACT compound cfmid results.csv
    - Approach 1 exp CE40 ENTACT compound cfmid results.csv
  + Column legend:
    - e0\_rank – Rank of ENTACT mixture compound score out of all candidate compound scores (against CE 10 CFMID predicted spectra)
    - e1\_rank – Rank of ENTACT mixture compound score out of all candidate compound scores (against CE 20 CFMID predicted spectra)
    - e2\_rank – Rank of ENTACT mixture compound score out of all candidate compound scores (against CE 40 CFMID predicted spectra)
    - total\_matches – Total number of candidate compounds for the given precursor mass
    - RANK\_E0 – Rank of ENTACT mixture compound score out of all candidate compound scores, only including candidate compounds with a formula matching ENTACT mixture compound (against CE 10 CFMID predicted spectra)
    - RANK\_E1 – Rank of ENTACT mixture compound score out of all candidate compound scores, only including candidate compounds with a formula matching ENTACT mixture compound (against CE 20 CFMID predicted spectra)
    - RANK\_E2 – Rank of ENTACT mixture compound score out of all candidate compound scores, only including candidate compounds with a formula matching ENTACT mixture compound (against CE 40 CFMID predicted spectra)
    - Formula\_matches – Total number of candidate compounds for the given precursor mass with a formula matching ENTACT mixture compound
    - Energy0 – CFMID score of ENTACT mixture compound (against CE 10 CFMID predicted spectrum)
    - Energy1 – CFMID score of ENTACT mixture compound (against CE 20 CFMID predicted spectrum)
    - Energy2 – CFMID score of ENTACT mixture compound (against CE 40 CFMID predicted spectrum)
    - MASS\_in\_MGF – Precursor mass from experimental MGF file
    - X\_percentile\_X\_bymass – Percentile value of ENTACT mixture compound score with respect to all other candidate compounds
    - X\_percentile\_X\_byformula – Percentile value of ENTACT mixture compound score with respect to all other candidate compounds, only including candidate compounds with a formula matching ENTACT mixture compound
    - X\_quot\_X\_bymass – Quotient value of ENTACT mixture compound score (where quotient value = [candidate compound CFMID score]/[Maximum CFMID score of all candidate comounds])
    - X\_quot\_X\_byform – Quotient value of ENTACT mixture compound score (where quotient value = [candidate compound CFMID score]/[Maximum CFMID score of all candidate comounds]), only including candidate compounds with a formula matching ENTACT mixture compound
* ENTACT v1 CFMID analysis (Approach 2).py
  + Excel files:
    - Approach 2 exp CE10 ENTACT compound cfmid results.csv
    - Approach 2 exp CE20 ENTACT compound cfmid results.csv
    - Approach 2 exp CE40 ENTACT compound cfmid results.csv
  + Column legend:
    - Rank\_bymass – Rank of ENTACT mixture compound score out of all candidate compound scores (where scores are summed across all three predicted CE spectra)
    - total\_matches – Total number of candidate compounds for the given precursor mass
    - RANK - Rank of ENTACT mixture compound score out of all candidate compound scores (where scores are summed across all three predicted CE spectra), only including candidate compounds with a formula matching ENTACT mixture compound
    - Formula\_matches - Total number of candidate compounds for the given precursor mass with a formula matching ENTACT mixture compound
    - SCORE - CFMID score of ENTACT mixture compound (where scores are summed across all three predicted CE spectra)
    - MASS\_in\_MGF – Precursor mass from experimental MGF file
    - Percentile\_by\_mass - Percentile value of ENTACT mixture compound score with respect to all other candidate compounds
    - Percentile\_by\_formula - Percentile value of ENTACT mixture compound score with respect to all other candidate compounds, only including candidate compounds with a formula matching ENTACT mixture compound
    - Quot\_by\_mass - Quotient value of ENTACT mixture compound score (where quotient value = [candidate compound CFMID score]/[Maximum CFMID score of all candidate comounds])
    - Quot\_by\_formula - Quotient value of ENTACT mixture compound score (where quotient value = [candidate compound CFMID score]/[Maximum CFMID score of all candidate comounds]), only including candidate compounds with a formula matching ENTACT mixture compound
* ENTACT v1 CFMID blinded analysis of data.py
  + This mode reads in three CFMID results files collected at different experimental CE levels (10, 20, 40) for one sample, and sums scores for candidate compounds across all three experimental CFMID results files. Used for Approach 3 processing.
  + Excel file:
    - ENTACT\_CFMID\_all\_mixtures-merged\_results\_rounded.csv
  + Column legend below:
    - DTXCID - DSSTox Chemical Identifier of candidate compound
    - MASS - monoisotopic mass of candidate compound
    - FORMULA – molecular formula of candidate compound
    - MATCHES – Number of candidate compounds with that specific molecular formula
    - Mixture – ENTACT mixture number for which the experimental MS2 spectra correspond to
    - Cfmid\_mode – Ionization mode of experimental spectrum
  + MASS\_in\_MGF – Experimental precursor mass from mgf file
  + SCORE – CFMID scores summed across both experimental CE levels (CE 10, 20, 40, when present for the specified precursor) and predicted CE levels (CE 10, 20, 40)
* ENTACT v1 CFMID analysis (Approach 3).py
  + Excel file
    - Approach 3 ENTACT compound cfmid results.csv
  + Column legend:
    - Percentile\_by\_mass - Percentile value of ENTACT mixture compound score with respect to all other candidate compounds
    - Percentile\_by\_formula - Percentile value of ENTACT mixture compound score with respect to all other candidate compounds, only including candidate compounds with a formula matching ENTACT mixture compound
    - Cfmid\_score\_mass\_quot - Quotient value of ENTACT mixture compound score (where quotient value = [candidate compound CFMID score]/[Maximum CFMID score of all candidate comounds])
    - Cfmid\_score\_form\_quot - Quotient value of ENTACT mixture compound score (where quotient value = [candidate compound CFMID score]/[Maximum CFMID score of all candidate comounds]), only including candidate compounds with a formula matching ENTACT mixture compound
    - Rank\_bymass – Rank of ENTACT mixture compound score out of all candidate compound scores (where scores are summed across all three predicted CE spectra)
    - total\_matches – Total number of candidate compounds for the given precursor mass
    - RANK - Rank of ENTACT mixture compound score out of all candidate compound scores (where scores are summed across all three predicted CE spectra), only including candidate compounds with a formula matching ENTACT mixture compound
    - Formula\_matches - Total number of candidate compounds for the given precursor mass with a formula matching ENTACT mixture compound
    - SCORE - CFMID score of ENTACT mixture compound (where scores are summed across all three predicted CE spectra)
    - MASS\_in\_MGF – Precursor mass from experimental MGF file
* In the case that an ENTACT mixture compound experimental spectrum did not have enough fragment matches to a predicted CFMID spectrum, a CFMID score of 0 is returned and the rank is set to -1 to indicate a poor match in the output files.