RESEARCH

MIMI: Molecular Isotope Mass Identifier for stable isotope-labeled Fourier transform ultra-high mass resolution data analysis

Michael A. Ochsenkühn [†1], Nabil Rahiman [†2], Shady A. Amin [1,2*] and Kristin C. Gunsalus [2,3*]

Abstract

Background: Ultra High Resolution (UHR) mass spectrometry systems with Fourier Transform Ion Cyclotron Resonance (FT-ICR) are often used to analyze the composition of complex mixtures of small molecules, such as metabolite samples for environmental, biological, or paleontological studies. The extremely high resolution of these systems enables simultaneous measurement of the exact masses of tens of thousands of molecular features and accurate determination of chemical formulas based on isotopic fine-structure ratios. To accelerate and streamline analysis of these datasets, automated solutions to rapidly characterize the molecular composition of unknown samples by comparison with known reference databases are needed.

Results: Here we present MIMI (Molecular Isotope Mass Identifier), a commandline tool to identify molecules present in complex samples using data from UHR-FT-ICR mass spectrometry. Given a database of known molecules and expected ratios of atomic isotopes in the sample, MIMI first computes chemical formulas, masses, and expected abundance for all possible isotopic variants in the database. Molecular formulas from publicly available databases and/or custom lists of molecules of interest can be used as reference data for comparison. By default MIMI is configured to use natural isotopic abundances, but it can easily accommodate different user-defined isotopic labeling ratios for any element(s). Candidate molecules are first identified in peak lists from UHR-FT-ICR mass spectrometry runs by comparing masses detected in the experimental data with precomputed expected masses for all entries in the reference database. Chemical formulas are then verified by searching for isotopic fine-structure matches and validated by counting measured and expected isotopes. We illustrate MIMI's utility using metabolite data from a cultured diatom sample isolated from sea water in the Arabian Gulf and spiked with ^{13}C-labeled internal standards.

Conclusions: We introduce a simple commandline tool, MIMI, that rapidly identifies chemicals present in samples of complex composition using UHR-FT-ICR mass spectrometry data. MIMI can compare measured data against both standard and customized molecular databases and can accommodate natural or user-specified isotope ratios. This software provides a convenient tool for simultaneous determination of natural and isotope-labeled compounds within the same sample, particularly for rapid characterization of complex mixtures of metabolites.

Keywords: FT-ICR; isotope labelling; metabolomics; mass spectrometry

Background

Fourier Transform Ion Cyclotron Resonance (FT-ICR) is a high-resolution mass spectrometry (MS) platform that uses ion cyclotron resonance frequencies and Fourier transformations to measure the mass-to-charge ratio (m/z) of ions in a sample. The resolving power of this technique lies in the sub-ppm range, making it possible to distinguish mass differences in the range of the mass of an electron [1, 2]. The platform's ability to provide accurate mass measurements, high resolving power, and high sensitivity make it particularly useful for analyzing complex mixtures of molecules [3]. It has been successfully applied to environmental, petroleum, and food/beverage samples, in addition to standard metabolomics applications [3, 4, 5]. Beyond provid-

*Correspondence: sa132@nyu.edu; kcg1@nyu.edu
[1] Marine Microbiomics Lab, Division of Biology, NYU Abu Dhabi, New York University, Abu Dhabi, United Arab Emirates
[3] Center for Genomics & System Biology, Department of Biology, New York University, New York, United States
Full list of author information is available at the end of the article

ing exact mass measurements, the ultra-high resolution of FT-ICR-MS allows the separation of molecular isotopes and identification of their chemical formulas. The expected isotopic abundance of atoms in a molecule (isotopic fine-structure) can then be used as a secondary determinant for correct chemical formula calculation [6, 7].

Moreover, enriching experiments with stable isotope labeling enables the investigation of elemental (carbon, oxygen, nitrogen, etc.) or molecule-specific changes in chemical and biological processes. This approach has been used in metabolomics, lipidomics, and proteomics to examine alterations in metabolic pathways, protein expression and turnover, flux analysis, and even intercellular interactions [8, 9, 10, 11, 12]. Isotope labeling is commonly used to trace contaminants in environmental studies, to study carbon and nitrogen metabolism in plants [13], and to characterize nitrogen fixation in microbes [14]; it has also been applied in medical [15, 16] and pharmacological studies [17]. The broad utility of isotope labeling and isotopic fine-structure mapping for molecular identification calls for automated tools to extract and analyze isotopic mass features from UHR mass spectrometry data.

To address the need for a versatile, isotope-aware analysis tool for UHR-FT-ICR mass spectrometry, we created the Molecular Isotope Mass Identifier (MIMI) tool. MIMI allows the user to specify any type and ratio of stable isotopes (e.g. ^{13}C, ^{15}N, ^{34}S) and thus can support a wide variety of studies and experimental designs, for example spiking natural samples with isotope-labeled compounds, uptake experiments with isotope-labeled metabolites, etc. MIMI first matches measured m/z values with expected values from one or more user-specified reference list(s) of chemical formulas, which may comprise data from publicly available metabolite databases and/or a list of specific molecules of interest. Since standard databases are incomplete and natural products discovery remains an active field, this flexibility is important for the study of model and non-model systems alike. To support the confidence in molecular identifications for exact mass spectra, MIMI then seeks to confirm the initial molecular assignments by comparing observed isotopic fine-structure patterns and relative peak abundances with expected patterns based on the reference data and given isotope ratios.

Implementation

MIMI is distributed as a Python package with a command-line interface. Code and documentation are available in a GitHub repository (see *Availability*).

A MIMI analysis essentially consists of comparing empirical mass spectra against expected masses for a list of reference compounds to identify the correct chemical formulas for molecules detected in MS datasets. Thus, MIMI will only identify molecular species in the data that are also present in the reference database rather than considering the entire chemical space. The MIMI design framework (Figure 1) separates the two main stages of the analysis into two distinct components: Pre-processing and mass analysis. These key features – modular architecture combined with reference database lookups – help to ensure fast processing speed, repeatability, and consistency.

Software architecture

The MIMI package contains only two main commands, corresponding to the two stages of the analysis (Figure 1). In the first stage (**Preprocessing**), `mimi_create_cache` builds a catalog of all possible isotopic variants for each chemical formula given in one or more list(s) of reference molecules, along with their masses and expected relative abundance, based on either natural isotope ratios or user-specified isotope composition. In the second stage (**Mass Analysis**), `mimi_mass_analysis` compares UHR mass spectra with reference data to identify matches and then uses their isotopic fine structure to verify molecular assignments.

The Preprocessing task is computationally intensive, but it needs to be performed only once for each combination of reference list and isotope composition. Separating the two stages thus dramatically reduces runtime when the same reference data are consulted repeatedly, for example, to run the same analysis for multiple samples and replicates or to test different error threshold parameters for peak matching.

Preprocessing

The `mimi_create_cache` command takes as input two types of files (Figure 2): one or more **molecular reference** file(s) in tabular (`.tsv`) format and a file specifying expected **atomic isotope ratios**, provided in `JSON` (`.json`) format. For each chemical formula in the reference list(s), MIMI combinatorially enumerates all possible isotopic variants and computes their molecular masses and expected relative abundances on the basis of the given isotope composition. These are written into a binary **cache** file (in `.pkl` format), which can be used to compare any subsequent MS data set expected to contain the same isotope composition with the same molecular reference list. Since each cache file represents a specific combination of chemical formulas and isotope ratios, a separate cache file must be created for each combination of these to be used for analysis.

Specifying a molecular database: The choice of reference molecules to be used for comparison with MS

spectra is specified with the `--database` (`-d`) parameter followed by one or more filenames. If more than one database file is included, MIMI will merge them during the Preprocessing stage. This can be useful, for example, when a user is particularly interested certain metabolites or other chemical compounds that are not present in a publicly available database. Note that since molecules with different chemical structures can share the same basic chemical formula, MIMI does not attempt to remove entries that might be redundant; it will, however, write to STDOUT the names and chemical formulas of entries found during parsing that share the same formula, in case the user would like to review these prior to deployment of the resulting cache file for mass analysis.

A properly formatted molecular reference file should include a header line containing at least three columns with the following names (in any order): **CF** (the chemical formula), **ID** (a unique identifier, such as from a publicly available database), and **Name** (a human-readable molecular name). Note that if additional columns are present (e.g. synonyms, description, associated enzymes, references, etc.), MIMI will ignore them when building the cache file, since only the chemical formula and ID are needed for the analysis (the Name will be retained for human-readable output).

For convenience, MIMI is packaged with a preformatted reference list from the publicly available Human Metabolome DataBase (HMDB v5.0) [18]. A helper script that can parse downloaded XML files from HMDB and reformat them for use with MIMI is also distributed with the package to accommodate future releases (see textitDocumentation for details). Additional reference files could be added as desired by the user.

Specifying isotope ratios: By default, MIMI will use atomic weights and natural isotope abundances from the National Institute of Standards and Technology (NIST), [19, 20] which are distributed with the MIMI package in JSON format. MIMI will always parse this file first as the basis for isotopic analysis. For any samples with stable isotope enrichment, it is necessary to explicitly specify new values for all elements with non-natural ratios. This allows MIMI to flexibly accommodate the presence of labeled organic compounds within biological metabolite samples. This is important since experimental studies increasingly employ stable isotope labeling with carbon (^{13}C), hydrogen (^{2}H), and/or nitrogen (^{15}N), or occasionally oxygen (^{17}O, ^{18}O) or sulfur (^{33}S, ^{34}S).

Users may override the default values for any element(s) using the `--label` (`-l`) option together with a user-defined `.json` file. An override file need specify only the new proportions of all isotopic variants for each labeled element(s); MIMI will simply update these entries in its preloaded list of natural isotope ratios. It is critical that the updated relative isotope abundances sum to 1.0; MIMI will perform a check for this and will throw an error if this check fails. For example, an experiment with 95% ^{13}C labeling corresponds to a ratio of 95:5 ^{13}C:^{12}C, so the override file should express the proportion of ^{13}C as 0.95 and that of ^{12}C as 0.05. An example override file specifying 95% ^{13}C labeling is included with the MIMI package as `C13_95.json` (see also Figure 2 for a partial illustration of this file).Note that if multiple isotopes are labeled in a single experiment (e.g. ^{13}C and ^{15}N together), the new values for all non-natural ratios should be included in a single override file, so that expected ratios of molecular variants are correctly specified for subsequent Mass Analysis.

Any non-natural labeling ratios supplied by the user should be either based on estimated ratios for the sample or, ideally, measured by isotope ratio mass spectroscopy (IRMS) to provide more accurate expected peak patterns for isotopic fine-structure comparison. Note that if incorrect isotope ratios are provided, or if an override file is omitted, initial peak matches with chemical formulas may still be detected, but the proportion of candidate assignments verified by fine-structure analysis will decrease.

Specifying ionization mode: To compute expected masses for the reference data, the ionization mode of the spectrometer must also be specified during the preprocessing step using the `--ion` (`-i`) parameter. Arguments are **positive** (`-i pos`) or **negative** (`-i neg`), for which MIMI will correspondingly add ($+H^+$) or subtract ($-H^+$) one proton mass of 1.007276467 from the computed mass. This version of MIMI only considers the most common adduct of a single ionization charge, since considering too many isotope variants causes a combinatorial explosion that would significantly increase computational complexity.

Specifying a cache file: The output of the Preprocessing stage is a cache file that contains theoretical masses for all isotopic variants of the molecules in the reference list(s), along with expected abundances based on the given isotope ratios. This information is stored as a serialized data object in a binary Python **pickle** file, which enables faster read/write times and uses less disk space than a text file. Unless otherwise specified by the user, the cache file will carry the same filename prefix as the input isotope ratio `.json` file, with the suffix `.pkl` (e.g. `C13_95.pkl`).

Mass Analysis

The `mimi_mass_analysis` command takes as input one or more 'pickled' cache files from the Preprocessing

step, along with one or more peak list(s) in tab-delimited `.acs` format (after calibration and peak picking, see *Materials and Methods*). The `--cache` (`-c`) option specifies the `.pkl` cache file(s) to be used for peak matching, while the `--sample` (`-s`) option specifies the sample(s) to be analyzed in a single run.

MIMI was designed to support comparisons of individual peak lists with multiple cache files and vice versa to accommodate scenarios that call for the combined analysis of one or more datasets using different expected isotope ratios. This is useful, for example, to identify molecules in a set of technical replicates of environmental or biological samples that have been spiked with isotope-labeled molecular standards for calibration. When multiple input files of either type are included in the same `mimi_mass_analysis` run, all results are written to a single output file to facilitate side-by-side comparisons, as shown in Figure 2.

Assignment of chemical formulas to mass spectral data is a two-part process: (1) peak lists are compared with the expected masses of reference molecules to identify matches with the most prevalent molecular variant (the monoisotopic mass), based on a given set of isotope ratios; and (2) preliminary assignments are verified by comparing observed vs. expected fine-structure patterns for isotopic variants of each candidate molecule.

For the initial peak matching, MIMI compares the sample and cache files using an efficient hash-based search algorithm. This enables the rapid identification of peak signals whose measured mass matches the theoretical monoisotopic mass of each reference molecule (given the expected isotope composition). The mass error threshold for the search window must be specified by the user in parts per million (ppm) using the `--ppm` (`-p`) option. The current resolution of UHR-FT-MS instruments is in the sub-ppm range, so an error window of ±0.5 ppm is a reasonable value for an initial MIMI run.

Once an exact match for the monoisotopic mass has been identified, MIMI seeks to verify the chemical formula assignment by comparing the observed fine-structure isotope pattern with that expected for the candidate compound. Comparisons are based on the precomputed masses and relative abundances of all molecular isotopes in a particular cache file and are performed using measured peak heights normalized with respect to the monoisotopic mass. The search window for matching isotopic variants is specified by the `-vp` option in ppm (where 'v' stands for 'validation').

Both the peak heights and relative abundances of the expected isotopes are considered at this stage. MIMI automatically applies two thresholds in its evaluation of fine-structure patterns. First, to account for experimental variation and integration errors, measured isotope peaks that are within ±30% of theoretical values are considered as 'validated'. Second, since background noise signals can give rise to false identifications, potential isotopic variants with a relative abundance below 10^{-5} of the monoisotopic mass are ignored. These cutoffs are somewhat arbitrary, but they are not expected to vary substantially between runs. Hence, the above values are included as defaults in a configuration file (`MIMI.config`) that is consulted by `mimi_mass_analysis` each time it is run, so that if the need arises these thresholds can be easily modified by the user.

Both the initial monoisotopic matching step and the subsequent fine-structure analysis can be tuned by varying the user-defined mass error tolerance thresholds (`-p` and `-vp`), which can affect the number of chemical formulas identified in the sample data. For example, exploring different `-vp` values can be useful because smaller peaks might present with broader peak shapes that can lead to shifts during peak picking. Users can test each dataset for the optimal thresholds by re-running the analysis across a range of values, as illustrated in Figure 3. This can be automated using a simple `bash` script, as outlined in the online *Documentation*. Other run parameters, including expected isotope ratios, can also be explored in a similar manner.

Output

The final product of a MIMI analysis is a tabular `.tsv` file containing a list of all the matched reference compounds and observed experimental data for the FT-ICR-MS data file(s) analyzed in the same `mimi_mass_analysis` run.

The output file, specified together with its relative file path using the `--output` (`-o`) parameter, will contain one or more rows of header labels followed by one row entry per matched chemical formula (Figure 2). As noted above, if the same formula is associated with more than one compound ID or name, all matches will be listed. The first several columns will contain molecular information from the reference database: chemical formula, database identifier, compound name, theoretical mass, and atom counts for individual elements (e.g. C, H, N, O, P, and S). Inclusion of atom counts facilitates the computation of various chemical properties that depend on atomic ratios, such as double bond equivalents (DBE), creation of van Krevelen diagrams, etc. This information will be followed by numerical data for each spectrum analyzed: measured mass (m/z), detected mass error (in ppm), peak intensity, and count of theoretical isotope peaks matched with measured data.

As noted above, when `mimi_mass_analysis` is run using multiple input files, the results will be concatenated in the same output file. If multiple `.acs` files are used, results for each dataset will be output side-by-side. Likewise, if more than one set of isotope ratios is used to analyze a single dataset, numerical data for each (`.pkl`) file will be presented in sequential columns under the same `.acs` file header (see *Documentation* for details). [NEED TO DETERMINE THE BEHAVIOR FOR NATURAL ABUNDANCE; IS IT ALWAYS PRESENT IN THE OUTPUT FILES?]

For example, say a spike-in of a set of known compounds labeled with 95% ^{13}C has been added as a control to a batch of samples or technical replicates. Since the fine-structure pattern of 95% ^{13}C-labeled molecules will differ from the pattern for natural abundances, it is desirable to compare each `.acs` file with both `natural_abundance.pkl` and `C13_95.pkl`. The output file for a combined run will report matches to entries in each cache file in separate groups for each dataset analyzed, enabling reference compounds in both the original sample and the spike-in to be identified independently for each dataset. This scenario is illustrated in the output file shown in Figure 2.

On the other hand, if each dataset to be analyzed contains a different spike-in, it will be inefficient (and undesirable) to analyze these together in the same run, since all `.acs` files will be compared against all `.pkl` files used as input for `mimi_mass_analysis`. For example, one sample may be labeled with ^{13}C only, while another maybe dual-labeled with both ^{13}C and ^{15}N. In such a case, it is preferable to run the analysis sequentially for each dataset, using a different combination of data and cache file in each run.

To enable reproducible analysis workflows, MIMI will also write a log file listing the full commands issued by `mimi_create_cache` and `mimi_analysis`, along with timestamps and a summary of the metadata. This provides a clear record of all input, cache, and output files together with parameter options used in each run. The name of the log file will match the prefix of the output filename concatenated with the suffix `.log`.

Full documentation for the MIMI package, including helper scripts for reformatting molecular database downloads (not described here), is available on the MIMI website (see *Documentation*).

Results and Discussion

To validate the tool's performance, we used MIMI to analyze FT-ICR-MS data for two technical replicates of a sample containing diatom dissolved organic matter (DOM) and spiked with a ^{13}C-labeled standard (see *Materials and Methods* for experimental details). The diatom strain, isolated from the Arabian Gulf, was cultured in axenic conditions, and de-salted diatom exudate was then spiked with IROA-IS internal molecular standards prior to mass spectrometry (IROA-IS is a yeast extract labeled with 95% ^{13}C, prepared using uniformly 95% ^{13}C-labeled glucose). FT-ICR-MS data were acquired using a time domain of 4 mega words across a mass range of 50-1000 m/z in negative ionization mode, resulting in 89,287 molecular features with a resolution of 750k at 110 m/z. The FT-ICR-MS test data files (`testdata1.asc, testdata2.asc`) are distributed with the MIMI package (see *Availability*).

Preprocessing with `mimi_create_cache` was performed using a database containing 14211 manually curated entries of metabolites with compound ID numbers (CO) from Kyoto Encyclopedia of Genes and Genomes (KEGG) as reference

The number of unique reference chemical formulas detected in the test data using these three mass error settings ranged from 637 to 2779 for unlabeled metabolites from diatom DOM, and from 118 to 925 for the 95% ^{13}C-labeled standard (Figure 3a). For both isotope compositions, we observed a 4- to 5-fold increase in the number of molecular features detected for an error threshold of `-p 0.5` vs. `-p 0.1` ppm, in line with the expected accuracy of the instrumentation. Consistent with the expectation that thousands of distinct molecules should be present in the unlabeled diatom sample, we detected a maximum of between 2181 and 2779 distinct molecular species in the test data. The expected ^{13}C-labeled IROA-IS spike-in composition of around 500-1000 KEGG compounds [21, **?**] also compares well with the 516 and 925 matched features at 0.5 and 1 ppm, respectively (Figure 3a).

In general, for FT-ICR-MS data with very high resolution (>250,000) deviations between measured and theoretical masses should be expected to be in the sub-ppm range. Depending on variations in instrumentation and acquisition parameters, standard FT-ICR-MS data should allow compound identification within a mass error of 0.5 ppm (after adequate acquisition and calibration and peak picking). Therefore, we choose this threshold for the exact mass match for further observations.

As peak shape might deteriorate for very small molecular peak features, we also explored the effect of varying the error rate for isotopic fine-structure feature comparisons. To illustrate isotope filtering and validation performance, we ran MIMI on the test data with the monoisotopic mass error fixed at `-p 0.5` and `-vp` settings of 0.1, 0.5 and 1 ppm (Figure 3b).

Between the two different isotope types, we should expect a higher number of detectable molecular isotopes for 95:5 ^{13}C:^{12}C-labeled molecules – since the relative abundance of all isotope peaks containing ^{12}C

will be higher than for the relative isotope abundance of ^{13}C 1.11% in natural isotope abundance molecules. For `-vp 0.1` ppm, we observed a very low number of isotopes due to insufficient instrument accuracy. For `-vp 0.5` and `-vp 1.0` ppm, an average of 3.1 and 5.3 isotopes, respectively, were matched with the natural isoforms. As expected, roughly double the number of isotopes – 5.9 and 10.3 – were matched to ^{13}C-labeled features at these settings.

Conclusion

Here we introduce MIMI (Molecular Isotope Mass Identifier), an open-source tool designed to simplify and accelerate the analysis of FT-ICR mass spectrometry data from both natural isotope abundance and isotope labeling experiments. MIMI enables efficient comparison of unlabeled and isotope-labeled chemical mixtures, offering flexibility through customizable reference databases and support for any combination of stable isotope ratios. By allowing users to input defined molecular databases and pre-process results for specific isotopes, MIMI significantly reduces computational time and hardware requirements. The tool processes multiple datasets simultaneously, generating a comprehensive result matrix that includes atom counts and isotope information in a single step, saving time as the number of samples increases. Although MIMI currently handles only MS1 data, future developments aim to integrate tandem mass spectrometry (MS2) to improve compound identification and structural interpretation. MIMI is well-suited for a wide range of experimental studies and serves as a foundation for advancing the analysis of ultra-high-resolution mass spectrometry data, particularly for biological and environmental research.

Materials and Methods

Diatom isolation and growth: *Asterionellopsis glacialis* strain A3 (CCMP3542) was isolated from the Arabian Gulf and kept axenic as described previously [22]. The cultures were maintained in f/2+Si [23] in semi-continuous batch cultures and incubated in Percival growth chambers (Percival Scientific, Perry, IA, USA) at 22°C, 130 µE $m^{-2}s^{-1}$ photosynthetic photon flux density (PPFD) and a 12h:12h light/dark cycle. Light flux was measured using a QSL-2100 PAR Sensor (Biospherical Instruments Inc., San Diego, CA, USA). Growth was monitored by measuring *in vivo* fluorescence using a 10-AU fluorometer (Turner Designs, San Jose, CA, USA).

Recovery of diatom exudate metabolites: To extract exudate molecules, the growth media of diatom culture at a RFU of ∼8 filtered through Whatman 0.2-µm polycarbonate membrane filters (Cytiva Life Sciences, Marlborough, MA, USA). Solid-phase extraction (SPE) was performed using Agilent PPL Bond Elut cartridges (1 g for 300 ml samples) (Agilent Technologies, Santa Clara, CA, USA). The columns were activated according to the manufacturer's instructions and subsequently used to remove and desalt organic molecules from the filtrates as previously described [24]. Finally, extracts were eluted with 1 ml and 5 ml 0.1% (v/v) formic acid in methanol and dried using a Savant SpeedVac SC210A evaporator (Thermo Fisher Scientific, Waltham, MA, USA), and stored at -20°C. For FT-ICR-MS analysis, the sample was redissolved in 400 µl of 90% MeOH and spiked with 15µl of IROA-IS ^{13}C-labeled internal standard dissolve according to the manufacturers recommendations in 1.2 ml ddH_2O(IROA Technologies, Ann Arbor, MI, USA).

FT-ICR-MS acquisition: High-resolution mass spectra were acquired on a Bruker solariX Fourier Transform Ion Cyclotron Resonance Fourier Transform Mass Spectrometer (FT-ICR-MS) (Bruker , Bellerica, MA, USA) equipped with a 7T superconducting magnet operated in the negative ionization mode. Spectra were acquired with a time domain of 4 mega words over a mass range of m/z 50 to 1000, with an optimal mass range from 200-600 m/z. Three hundred scans were accumulated for each sample replicate at a sample injection rate of 1 µl/min. The following parameters were set for molecule detection in the m/z 100 to 1000 range: Source: ESI, negative mode, Capillary: 4500 V, End Plate Offset: -700V, Nebulizer pressure: 2 bar, Dry gas flow: 10 L/min, Dry temperature: 220°C. Ion-trap: Source Optics, Capillary Exit: -200V, Deflector Plate: -220V, Funnel1: -150V, Skimmer: -15V, Funnel RF App: 150Vpp; Octopole Frequency: 5 MHz, RF Amplitude: 350 Vpp, Collison Cell: Collision Voltage: -10V, DCextract Bias: -0.6V, RF Frequency: 2 Mhz, Collision RF Amplitude: 1500 Vpp, Transfer Optics: TOF: 0.5 ms, Frequency: 4 MHz, RF Amplitude 350 Vpp. Analyzer: ParaCell: Transfer Exit Lens: 23.0 V, Analyzer Entrance: 10.0 V, Side kick: 0.0 V, Side kick Offset: 1.5 V, Front trap plate: -2.0 V, Back Trap plate: -2.0 V, Back trap Quench: 30 V, Sweep excitation: 21%; Shimming DC Bias: 0°: -1.468, 180°: -1.532, 90°: -1.420, 270°:-1.584, ICR Fills: 1.

DataAnalysis 5.0 calibration and feature picking: The FT-ICR mass spectra were internally calibrated on primary metabolites (amino acids, fatty acids, and organic acids) using DataAnalysis Version 5.0 (Bruker , Bellerica, MA, USA). Peak alignment was performed with maximum error thresholds of 0.01 ppm and with a cut-off signal-to-noise (S/N) ratio of 4. The resulting peak table was exported to peak lists in tab-delimited `.asc` format.

Software

For this study, MIMI V1.0 was run using Python, Version 3.12.1 (https://www.python.org/, Python Software Foundation) running under Anaconda (Anaconda

Inc., Austin, TX, USA, 2023). The analysis has been performed on both MAC and PC systems.

Declarations

Availability and requirements

MIMI is distributed as an open-source Python software package under an NYU Non-Commercial License. The MIMI codebase, along with reported example files, test data and other associated files, can be found on GitHub at https://github.com/GunsalusPiano/mass_spectrometry_tool/.
Full documentation for installation and usage of the software may be found at https://gunsaluspiano.github.io/mass_spectrometry_tool/.

- **Project name**: MIMI (Molecular Isotope Mass Identifier)
- **Software version**: MIMI v1.0
- **Code repository**: https://github.com/NYUAD-Core-Bioinformatics/MIMI
- **Documentation**: https://corebioinf.abudhabi.nyu.edu/MIMI
- **Bioconda package**: ???
- **Operating systems:** Linux, MacOS, Windows.
- **Programming language:** Python 3.
- **License:** NYU non-commercial research license, free for academic use.
- **Commercial use:** License required; see Github repository for contact details.

Availability of data and materials

All test datasets, configuration files, and associated materials discussed in the manuscript are available in the MIMI GitHub repository.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

MAO conceived the project; NR developed the algorithms and wrote the code; SA and KCG provided supervision and guidance; and all authors contributed to writing the manuscript.

Funding

This work was supported by a Gordon and Betty Moore Foundation award to SAA (GBMF9335, https://doi.org/10.37807/GBMF9335), NYU Abu Dhabi award to SAA (AD179), and by Tamkeen under the NYU Abu Dhabi Research Institute Award for the NYUAD Center for Genomics and Systems Biology to KCG (ADHPG-CGSB).

Acknowledgements

The authors would like to acknowledge the NYU Abu Dhabi Core Technology Platforms facilities for access to the FT-ICR-MS. [KCG to Nabil: any HPC usage?]

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Author details

[1] Marine Microbiomics Lab, Division of Biology, NYU Abu Dhabi, New York University, Abu Dhabi, United Arab Emirates. [2] NYU Abu Dhabi Center for Genomics & System Biology, New York University, Abu Dhabi, United Arab Emirates. [3] Center for Genomics & System Biology, Department of Biology, New York University, New York, United States.

References

1. Nikolaev, E.N.: Some notes about FT ICR mass spectrometry **377**, 421–431. doi:10.1016/j.ijms.2014.07.051. Accessed 2024-02-15
2. Marshall, A.G., Hendrickson, C.L., Jackson, G.S.: Fourier transform ion cyclotron resonance mass spectrometry: A primer **17**(1), 1–35. doi:10.1002/(SICI)1098-2787(1998)17:1¡1::AID-MAS1¿3.0.CO;2-K
3. Ghaste, M., Mistrik, R., Shulaev, V.: Applications of fourier transform ion cyclotron resonance (FT-ICR) and orbitrap based high resolution mass spectrometry in metabolomics and lipidomics **17**(6), 816. doi:10.3390/ijms17060816
4. Kanawati, B., Wanczek, K.P., Schmitt-Kopplin, P.: Data processing and automation in fourier transform mass spectrometry. In: Fundamentals and Applications of Fourier Transform Mass Spectrometry, pp. 133–185. doi:10.1016/B978-0-12-814013-0.00006-5. http://www.scopus.com/inward/record.url?scp=8508158768
5. Rodgers, R.P., Marshall, A.G.: Petroleomics: Advanced characterization of petroleum-derived materials by fourier transform ion cyclotron resonance mass spectrometry (FT-ICR MS). In: Mullins, O.C., Sheu, E.Y., Hammami, A., Marshall, A.G. (eds.) Asphaltenes, Heavy Oils, and Petroleomics, pp. 63–93. Springer. doi:10.1007/0-387-68903-6$_3$.
6. Miladinović, S.M., Kozhinov, A.N., Gorshkov, M.V., Tsybin, Y.O.: On the utility of isotopic fine structure mass spectrometry in protein identification **84**(9), 4042–4051. doi:10.1021/ac2034584
7. Nagao, T., Yukihira, D., Fujimura, Y., Saito, K., Takahashi, K., Miura, D., Wariishi, H.: Power of isotopic fine structure for unambiguous determination of metabolite elemental compositions: *In silico* evaluation and metabolomic application **813**, 70–76. doi:10.1016/j.aca.2014.01.032. Accessed 2024-02-15
8. Zhang, R., Chen, B., Zhang, H., Tu, L., Luan, T.: Stable isotope-based metabolic flux analysis: A robust tool for revealing toxicity pathways of emerging contaminants **159**, 116909. doi:10.1016/j.trac.2022.116909. Accessed 2024-02-15
9. Qian, Z.-Y., Ma, J., Sun, C.-L., Li, Z.-G., Xian, Q.-M., Gong, T.-T., Xu, B.: Using stable isotope labeling to study the nitrogen metabolism in anabaena flos-aquae growth and anatoxin biosynthesis **127**, 223–229. doi:10.1016/j.watres.2017.09.060
10. Chokkathukalam, A., Kim, D.-H., Barrett, M.P., Breitling, R., Creek, D.J.: Stable isotope-labeling studies in metabolomics: new insights into structure and dynamics of metabolic networks **6**(4), 511–524. doi:10.4155/bio.13.348. Accessed 2024-02-15
11. Shih, L.-M., Tang, H.-Y., Lynn, K.-S., Huang, C.-Y., Ho, H.-Y., Cheng, M.-L.: Stable isotope-labeled lipidomics to unravel the heterogeneous development lipotoxicity **23**(11), 2862. doi:10.3390/molecules23112862
12. Gevaert, K., Impens, F., Ghesquière, B., Van Damme, P., Lambrechts, A., Vandekerckhove, J.: Stable isotopic labeling in proteomics **8**(23), 4873–4885. doi:10.1002/pmic.200800421
13. Maia, M., Figueiredo, A., Cordeiro, C., Sousa Silva, M.: FT-ICR-MS-based metabolomics: A deep dive into plant metabolism **42**(5), 1535–1556. doi:10.1002/mas.21731
14. Angel, R., Panhölzl, C., Gabriel, R., Herbold, C., Wanek, W., Richter, A., Eichorst, S.A., Woebken, D.: Application of stable-isotope labelling techniques for the detection of active diazotrophs **20**(1), 44–61. doi:10.1111/1462-2920.13954. Accessed 2024-02-15
15. Mutlib, A.E.: Application of stable isotope-labeled compounds in metabolism and in metabolism-mediated toxicity studies **21**(9), 1672–1689. doi:10.1021/tx800139z
16. Knapp, D.R., Gaffney, T.E.: Use of stable isotopes in pharmacology–clinical pharmacology **13**(3), 307–316. doi:10.1002/cpt1972133307
17. Schellekens, R.C.A., Stellaard, F., Woerdenbag, H.J., Frijlink, H.W., Kosterink, J.G.W.: Applications of stable isotopes in clinical pharmacology **72**(6), 879–897. doi:10.1111/j.1365-2125.2011.04071.x. Accessed 2024-02-15
18. Wishart, D.S., Guo, A., Oler, E., Wang, F., Anjum, A., Peters, H., Dizon, R., Sayeeda, Z., Tian, S., Lee, B.L., Berjanskii, M., Mah, R., Yamamoto, M., Jovel, J., Torres-Calzada, C., Hiebert-Giesbrecht, M., Lui, V.W., Varshavi, D., Varshavi, D., Allen, D., Arndt, D., Khetarpal, N., Sivakumaran, A., Harford, K., Sanford, S., Yee, K., Cao, X., Budinski, Z., Liigand, J., Zhang, L., Zheng, J., Mandal, R., Karu, N., Dambrova, M., Schiöth, H.B., Greiner, R., Gautam, V.: HMDB 5.0: the Human Metabolome Database for 2022. Nucleic Acids Research **50**(D1), 622–631 (2022). doi:10.1093/nar/gkab1062
19. Coursey, J.S., Schwab, D.J., Tsai, J.J., Dragoset, R.A.: Atomic Weights and Isotopic Compositions (version 4.1) (2015)
20. Meija, J., Coplen, T.B., Berglund, M., Brand, W.A., Bièvre, P.D., Gröning, M., Holden, N.E., Irrgeher, J., Loss, R.D., Walczyk, T., Prohaska, T.: Atomic weights of the elements 2013 (IUPAC technical report) **88**(3), 265–291. doi:10.1515/pac-2015-0305. Publisher: De Gruyter. Accessed 2024-02-15
21. Beecher, C., de Jong, F.: TruQuant Yeast Extract Workflow KIT

22. Behringer, G., Ochsenkühn, M.A., Fei, C., Fanning, J., Koester, J.A., Amin, S.A.: Bacterial communities of diatoms display strong conservation across strains and time **9**, 659. doi:10.3389/fmicb.2018.00659
23. Guillard, R.R., Ryther, J.H.: Studies of marine planktonic diatoms. i. cyclotella nana hustedt, and detonula confervacea (cleve) gran **8**, 229–239. doi:10.1139/m62-029
24. Ochsenkühn, M.A., Schmitt-Kopplin, P., Harir, M., Amin, S.A.: Coral metabolite gradients affect microbial community structures and act as a disease cue **1**(1), 1–10. doi:10.1038/s42003-018-0189-1. Number: 1 Publisher: Nature Publishing Group. Accessed 2024-02-15

Figures

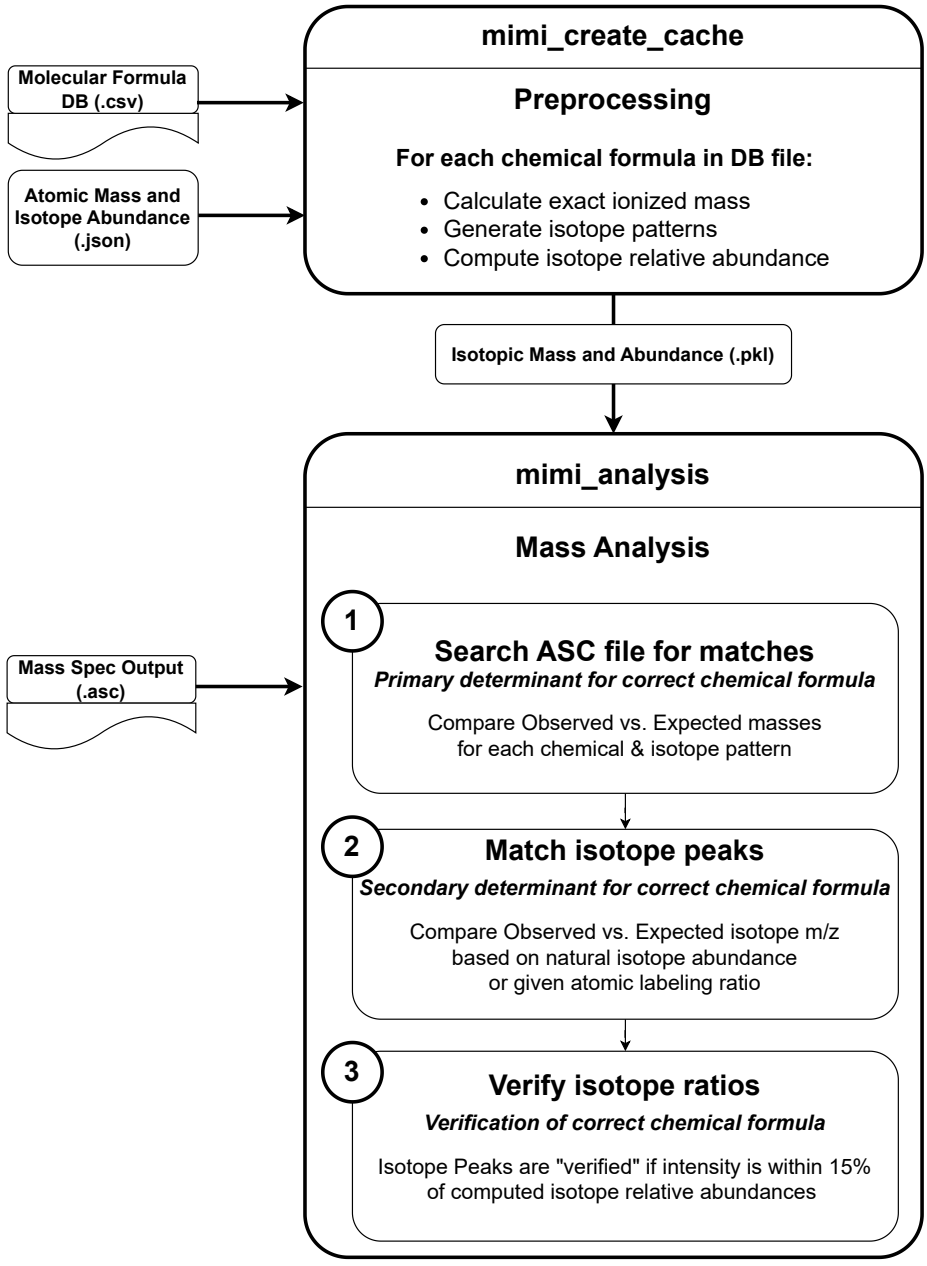

Figure 1 MIMI Architecture: Flowchart describing Preprocessing and Mass Analysis modules.
[KCG: I am tweaking this figure to reflect multiple file inputs and minor text adjustments.]

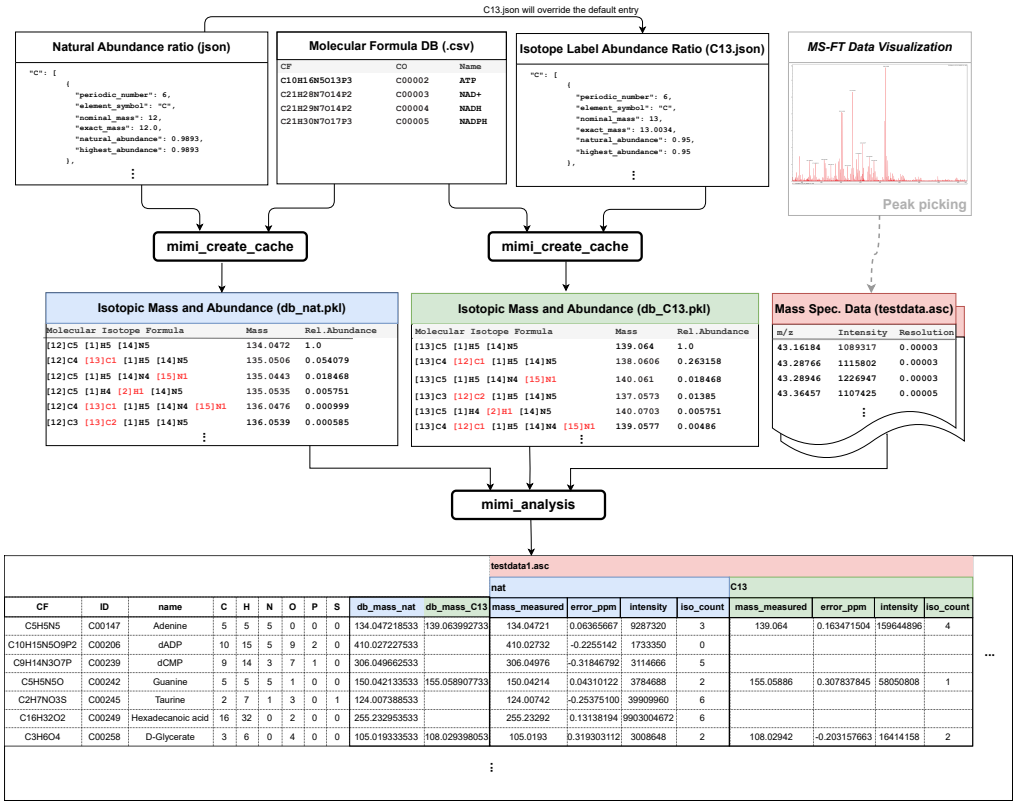

Figure 2 MIMI Workflow: A workflow visualizing the data structure and dependencies for the Preprocessing (`mimi_create_cache`) and Mass Analysis (`mimi_analysis`) steps, along with the expected output data.

[KCG: add note about column headers?; perhaps also add explanatory note about the isotope file override here? We need to fix the figure, there are a few things I think we could improve.]

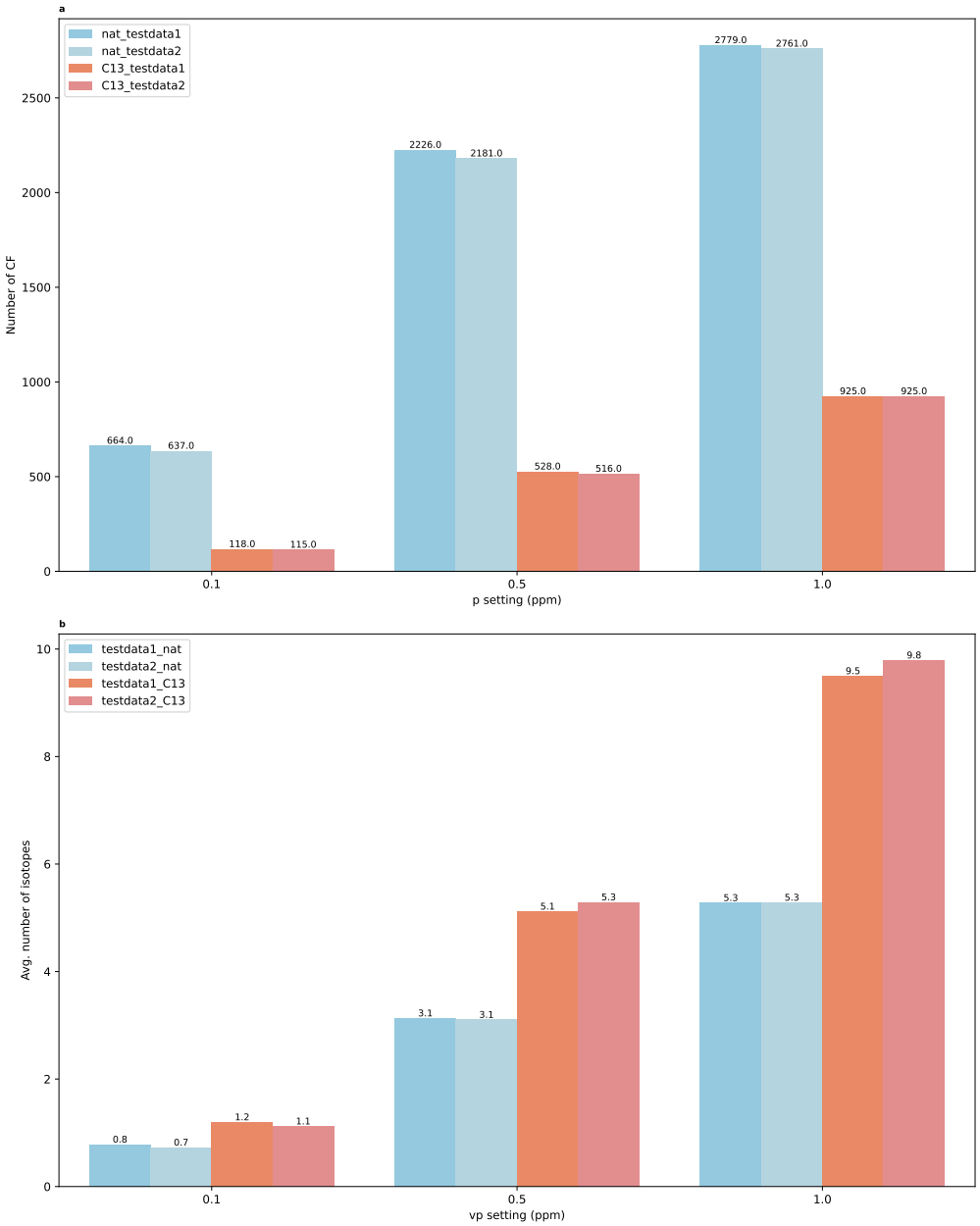

Figure 3 Comparison of calculated molecular features for natural isotopic abundance (blue) and ^{13}C labelled (orange) chemical formulas from IROA-IS with varying -p at 0.1, 0.5, 1. b) Number of average isotopes found for compounds found with a -p setting of 0.5 at varying -vp values of 0.1, 0.5, 1 ppm, and c) average number of verified isotopes.

[KCG: First, the labels on this graph are WAAAAY to small; and, there is no legend for part a but there is a legend for a non-existent part c.]