gcms_data_analysis: A Python package for automated analysis of GC-MS data

20 February 2024

Summary

The lack of open-source tools to automate the analysis of large datasets from Gas Chromatography coupled with Mass Spectrometry (GC-MS) of biofuels results in time-consuming manual analyses of such data that employ sub-optimal methodologies, are difficult to replicate, and present an increased risk of error. We developed a Python code to automate GC-MS data analysis of complex, heterogeneous organic mixtures. The code retrieves properties for each identified chemical from PubChem, applies calibration and semi-calibration using Tanimoto similarity index and molecular weight similarity, splits each identified compound into its functional groups, and handles both derivatized and non-derivatized samples. Replicates of the same samples are automatically combined to compute averages and standard deviations. The outputs include single and multiple sample reports with area, concentrations and yield on feedstock basis and aggregated reports based on the cumulative fraction of each functional group in each analyzed sample. This tool reduces GC-MS analysis time from hours/days to few seconds, avoiding human-errors and promoting standardization and best practices for GC-MS data handling.

Statement of need

Gas Chromatography-Mass Spectrometry (GC-MS) is routinely used to identify and quantify organic mixtures. In particular, GC-MS analysis of liquid biofuels enables assessment of the composition and therefore quality and value of these sustainable fuels (Lu et al. 2017; Sharma et al. 2020; Sugumaran et al. 2017). The detailed compositional analysis from GC-MS of bio-oil can reveal the chemical reactions underpinning biofuel production processes, thus enabling the optimization of process conditions based on feedstock composition (Grams 2020; Sudibyo, Pecchi, and Tester 2022; Wang et al. 2020).

Performing qualitative GC-MS analysis of bio-oil samples is relatively straightforward; bio-oil is routinely extracted using an organic solvent and compounds identified by spectral library matches (Heracleous et al. 2022; Kostyukevich et al. 2019; Zhu et al. 2022). In theory, the quantification of compounds via GC-MS is rather simple; chromatogram area is linearly related to concentration of each analyte. In practice, quantifying the vast numbers of compounds detected in heterogeneous bio-oils requires the construction of a large calibration dataset (Kohansal et al. 2022; Panisko et al. 2015; Villadsen et al. 2012). Depending on the GC-MS column, compounds present, and their concentrations, derivatization may be necessary to improve identification, reproducibility and therefore final quantification of species present (Leonardis et al. 2013; Madsen et al. 2016; Wang et al. 2020). Derivatization requires its own calibration to quantify derivatized peaks, sometimes at different operating conditions than the non-derivatized analysis. These experimental challenges are compounded by the handling and analysis of the large amount of data that GC-MS analyses produce. A single bio-oil sample can contain hundreds of compounds (Haider, Castello, and Rosendahl 2020; Han et al. 2021). A typical experimental campaign involves dozens of samples (hopefully with at least a modicum of replication). Thus, the analysis of such datasets is onerous; the application of a calibration curve to quantify each individual component in a bio-oil, and the translation from derivatized results to the original bio-oil components, both add downstream computational tasks. Manual handling of such large datasets introduces the risk of human error.

In general, most GC-MS instruments provide solutions for the identification of compounds through fragmentation pattern matching by means of proprietary software, and several open-source options are available for the task (O'Callaghan et al. 2012). For each sample, these programs produce a semi-quantitative table with the identified compounds and their measured chromatogram area, which indicates the intensity of the signal for each compound. If a user has developed a calibration curve for a particular compound, the instrument's software will (usually) report the concentration of the component in the sample vial. For derivatized samples, the derivatized version of the compound is usually returned (since the derivatized molecule enters the machine), which requires an additional step to translate the derivatized versions into their non-derivatized (original) compound. Most instrument companies' proprietary software handle calibrations, but their use can be counterintuitive and/or difficult to implement in additional automatic data analysis routines. Both from our own experience and conversations with researchers around the world, calibrations are often applied manually by researchers, especially when a semi-calibration mode is adopted to estimate the concentration of non-calibrated compounds based on similar compounds' calibration curves.

Another bottleneck in the analysis of GC-MS data is the classification of identified compounds based on their functional group(s). This is especially useful for heterogeneous mixtures such as bio-oils and other thermochemical and chemical processes where aggregated compositional results (e.g., the fraction of compounds having a given functional group present in each sample) can be used to derive mechanistic conclusions (Heracleous et al. 2022; Sudibyo, Pecchi, and Tester

2022; Yang et al. 2018; Zhu et al. 2022). While performing these tasks manually is possible and has led to valuable results, automation could save time and minimize human error, while also increasing transparency in the methodological approach, specifically in how functional groups are attributed to each compound. Instruments' proprietary software is not always user-friendly or intuitive, and often requires extensive training, especially when used for quantitation of non-calibrated compounds. Such proprietary software does not allow for automation of the full GC-MS data analysis pipeline, usually limiting the automation at the single sample analysis. As such, a new tool for researchers that is open-access and incorporates multiple automated analysis pathways could save researchers' time and promote a higher level of standard best practices across GC-MS users, especially in the biofuels field.

To address this need, we designed an open-source Python tool to fully automate the handling of multiple GC-MS datasets simultaneously. The tool relies on the Python package PubChemPy (Swain 2017) to access the PubChem website (Kim et al. 2023) and build its own database with all identified chemicals and their relevant chemical properties. It also performs functional group fragmentation using each compound's SMILES (Weininger 1988) retrieved from PubChem, employing the automatic fragmentation algorithm developed by (Müller 2019) to split each molecule into its functional groups and then assign to each group its mass fraction. Functional groups can also be specified by the user using their SMARTS codes (Daylight Chemical Information Systems, n.d.). The code can apply calibrations for quantifying components present in a sample, with an semi-calibration option based on Tanimoto similarity with tunable thresholds (Bajusz, Rácz, and Héberger 2015; Chen and Reynolds 2002). For derivatized samples, the procedure is unchanged except that the non-derivatized form of each identified compound is used to retrieve chemical information so that non-derivatized and derivatized samples can be directly compared.

Besides producing single sample reports, the tool produces comprehensive reports that include all compounds across a series of samples as well as aggregated reports for all samples based on functional group mass fractions. The code also provides plotting functions for the aggregated reports to visualize the results and enable a preliminary investigation of their statistical significance.

Automatic fragmentation and aggregation by functional group

Generally, when discussing chemical differences among samples (for example, with the aim of assessing variation in process mechanism when different process parameters are employed), describing the differences in terms of single chemical compounds across said samples may not be a feasible approach due to the large number of compounds and their low concentrations and therefore low relative importance. A typical strategy is to aggregate compounds based on chemical families, usually based on their functional group, either by count or concentrationconcentration (Castello, Haider, and Rosendahl 2019; Heracleous et al. 2022; Sudibyo, Pecchi, and Tester 2022; Zhu et al. 2022). An effective approach (though extremely time consuming if performed manually) is to calculate the mass fraction of each functional group present in the compound and use these mass fractions to split the area or the concentration of each molecule into weighted averages of the different functional groups present, and then compute the aggregated value for all compounds present in the sample. This has the advantage of fully accounting for all functional groups in a given sample. The aggregated concentration of each functional group (C_{fg}) in the sample is obtained as the sum over all n identified compounds in the sample of the concentration of the compound (C_i) multiplied by the mass fraction of the considered functional group in the compound itself ($mf_{fg,i}$). This equally applies to area, concentration, and yield:

$$C_{fg} = \sum_{i=1}^{n} C_i \cdot mf_{fg,i}$$

This latter approach is automated in the present tool. The fragmentation into functional groups relies on the fragmentation algorithm developed by (Müller 2019), available on GitHub.

Semi-calibration based on Tanimoto and molecular weight similarity

Since calibrating with pure standards for each identified compound in complex samples such as heterogeneous bio-oil is likely infeasible due to the large number of identified compounds, most bio-oil discussions rely on relative concentrations based on identified peak areas. In other cases, GC-MS are calibrated for some of the identified compounds, and authors adopt strategies to infer calibrations for compounds without existing calibration Hubble and Goldfarb (2021). In general, these methods are often manually implemented and revolve around the use of some sort of nearest neighbor method, where for each compound without a calibration curve, the calibration available for the "closest" compound is used instead. An effective approach, so far unexplored in the biofuel GC-MS literature but common in cheminformatics (Butina 1999), is to use molecular fingerprints (encodings of the molecular structure) to compute similarity indices to select the most similar calibrated compound. A popular choice in the cheminformatics literature is the Tanimoto similarity index (Bajusz, Rácz, and Héberger 2015; Chen and Reynolds 2002). The present code allows for semi-calibration (if the option is selected); for compounds without a calibration curve, their Tanimoto similarity with all calibrated compounds is evaluated. The calibrated compound with the highest similarity is selected (if more compounds share the same similarity, as happens for compounds of the same class, the compound with the closest molecular weight is selected). The Python package rdkit (al., n.d.) is used to convert canonical SMILES into molecular fingerprints and to compute the Tanimoto similarity among compounds.

Error associated with Tanimoto thresholds

The error associated with the use of the semi-calibration approach (for example, assuming use of the calibration curve of compound c1 to estimate the concentration of compound c2) can be evaluated, if both c1 and c2 calibration curves are known, as the average error between the calibration curves of those two compounds. This error equals the error that would come from using the calibration of c1 for estimating the concentration of c2, should c2 not be available. The average error can be computed using the following equation, where cal_{c1} and cal_{c2} are the calibration curves for c1 and c2 obtained by the linear interpolation of runs at known concentrations:

$$\text{Average error [\%]} = \overline{\left(\frac{|cal_{c1} - cal_{c2}|}{cal_{c1}}\right)} \times 100$$

We assessed this error for all combinations of calibrated compounds calibration dataset available in our GC-MS. The dataset comprises 89 compounds for which more than 4 points calibration (up to 6 points) is available in the form of mg/L vs detected area; this results in 3827 combinations. For each combination of compounds, the Tanimoto similarity and the molecular weight difference is also computed. Figure Figure 1 plots the average error as a function of the Tanimoto similarity of compounds and their molecular weight difference. The error decreases with the increasing Tanimoto similarity, while there seems to be no marked effect of molecular weight difference. Selecting a Tanimoto similarity threshold of 0.4 minimizes the risk of errors that are above one order of magnitude, while a similarity of 0.7 avoids this almost entirely (at least in our dataset). The percentage error threshold of 100% may seem unreasonably high, however the alternative would be to simply ignore all the compounds for which no calibration is available which also implies a large data loss. The selection of a Tanimoto similarity threshold that avoids unrealistic overestimation of concentrations (underestimations are less of a concern, since the alternative would be to ignore the compound entirely) improves the quality of GC-MS results. A similarity threshold of 0.4 was arbitrarily chosen as the default threshold in the code based on these results, with a molecular weight difference a threshold of 100 atomic mass units set as the default.

Acknowledgements

This work was supported by the National Science Foundation CBET under grant numbers 1933071 and 2031710. The authors thank James Li Adair for the constructive criticism on the manuscript; Alessandro Cascioli for the help

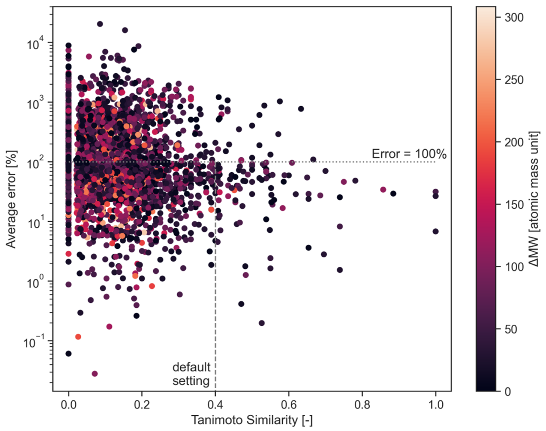

Figure 1: Average error between calibration curves of combination of compounds as a function of their Tanimoto similarity. The molecular weight difference of the compounds is reported in the colormap. The horizontal like indicates 100% error, the vertical line reports the default similarity threshold adopted in the code.

on Python package structuring; and Alessandro Cascioli, James Li Adair, and Madeline Karod for being early testers of the code.

References

Ahn, Ji-Won, Sudhir Kumar Pandey, and Ki-Hyun Kim. 2011. "Comparison of GC-MS Calibration Properties of Volatile Organic Compounds and Relative Quantification Without Calibration Standards." *Journal of Chromatographic Science* 49 (1): 19–28. https://doi.org/10.1093/chrsci/49.1.19.

al., Greg Landrum; Paolo Tosco; Brian Kelley; et. n.d. "RDKit: Open-Source Cheminformatics." https://www.rdkit.org.

Bajusz, Dávid, Anita Rácz, and Károly Héberger. 2015. "Why Is Tanimoto Index an Appropriate Choice for Fingerprint-Based Similarity Calculations?" *Journal of Cheminformatics* 7 (1): 20. https://doi.org/10.1186/s13321-015-0069-3.

Butina, Darko. 1999. "Unsupervised Data Base Clustering Based on Daylight's Fingerprint and Tanimoto Similarity: A Fast and Automated Way to Cluster Small and Large Data Sets." *Journal of Chemical Information and Computer Sciences* 39 (4): 747–50. https://doi.org/10.1021/ci9803381.

Castello, Daniele, Muhammad Salman Haider, and Lasse Aistrup Rosendahl. 2019. "Catalytic Upgrading of Hydrothermal Liquefaction Biocrudes: Different Challenges for Different Feedstocks." *Renewable Energy* 141 (October): 420–30. https://doi.org/10.1016/j.renene.2019.04.003.

Chen, Xin, and Charles H. Reynolds. 2002. "Performance of Similarity Measures in 2D Fragment-Based Similarity Searching: Comparison of Structural Descriptors and Similarity Coefficients." *Journal of Chemical Information and Computer Sciences* 42 (6): 1407–14. https://doi.org/10.1021/ci025531g.

Daylight Chemical Information Systems, Inc. n.d. "SMARTS - a Language for Describing Molecular Patterns." https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html.

Grams, Jacek. 2020. "Chromatographic Analysis of Bio-Oil Formed in Fast Pyrolysis of Lignocellulosic Biomass." *Reviews in Analytical Chemistry* 39 (1): 65–77. https://doi.org/10.1515/revac-2020-0108.

Haider, Muhammad Salman, Daniele Castello, and Lasse Aistrup Rosendahl. 2020. "Two-Stage Catalytic Hydrotreatment of Highly Nitrogenous Biocrude from Continuous Hydrothermal Liquefaction: A Rational Design of the Stabilization Stage." *Biomass and Bioenergy* 139 (August): 105658. https://doi.org/10.1016/j.biombioe.2020.105658.

Han, Jiahui, Xing Li, Shengyan Kong, Guang Xian, Hualong Li, Xun Li, Jie Li, et al. 2021. "Characterization of Column Chromatography Separated Bio-Oil Obtained from Hydrothermal Liquefaction of Spirulina." *Fuel* 297 (August): 120695. https://doi.org/10.1016/j.fuel.2021.120695.

Heracleous, Eleni, Michalis Vassou, Angelos A. Lappas, Julie Katerine Rodriguez, Stefano Chiaberge, and Daniele Bianchi. 2022. "Understanding the Upgrading of Sewage Sludge-Derived Hydrothermal Liquefaction Biocrude via Advanced Characterization." *Energy & Fuels* 36 (19): 12010–20. https://doi.org/10.1021/acs.energyfuels.2c01746.

Hubble, Andrew H., and Jillian L. Goldfarb. 2021. "Synergistic Effects of Biomass Building Blocks on Pyrolysis Gas and Bio-Oil Formation." *Journal of Analytical and Applied Pyrolysis* 156 (June): 105100. https://doi.org/10.1016/j.jaap.2021.105100.

Kim, Sunghwan, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, et al. 2023. "PubChem 2023 Update." *Nucleic Acids Research* 51 (D1): D1373–80. https://doi.org/10.1093/nar/gkac956.

Kohansal, Komeil, Kamaldeep Sharma, Muhammad Salman Haider, Saqib Sohail Toor, Daniele Castello, Lasse Aistrup Rosendahl, Joscha Zimmermann, and Thomas Helmer Pedersen. 2022. "Hydrotreating of Bio-Crude Obtained from Hydrothermal Liquefaction of Biopulp: Effects of Aqueous Phase Recirculation on the Hydrotreated Oil." *Sustainable Energy & Fuels* 6 (11): 2805–22. https://doi.org/10.1039/D2SE00399F.

Kostyukevich, Yury, Mihail Vlaskin, Alexander Zherebker, Anatoly Grigorenko, Ludmila Borisova, and Eugene Nikolaev. 2019. "High-Resolution Mass Spectrometry Study of the Bio-Oil Samples Produced by Thermal Liquefaction of Microalgae in Different Solvents." *Journal of The American Society for Mass Spectrometry* 30 (4): 605–14. https://doi.org/10.1007/s13361-018-02128-9.

Leonardis, Irene, Stefano Chiaberge, Tiziana Fiorani, Silvia Spera, Ezio Battistel, Aldo Bosetti, Pietro Cesti, Samantha Reale, and Francesco De Angelis. 2013. "Characterization of Bio-Oil from Hydrothermal Liquefaction of Organic Waste by NMR Spectroscopy and FTICR Mass Spectrometry." *ChemSusChem* 6 (1): 160–67. https://doi.org/10.1002/cssc.201200314.

Lu, Yao, Guo-Sheng Li, Yong-Chao Lu, Xing Fan, and Xian-Yong Wei. 2017. “Analytical Strategies Involved in the Detailed Componential Characterization of Biooil Produced from Lignocellulosic Biomass.” *International Journal of Analytical Chemistry* 2017: 9298523. https://doi.org/10.1155/2017/9298523.

Madsen, René Bjerregaard, Mads Mørk Jensen, Anders Juul Mørup, Kasper Houlberg, Per Sigaard Christensen, Maika Klemmer, Jacob Becker, Bo Brummerstedt Iversen, and Marianne Glasius. 2016. “Using Design of Experiments to Optimize Derivatization with Methyl Chloroformate for Quantitative Analysis of the Aqueous Phase from Hydrothermal Liquefaction of Biomass.” *Analytical and Bioanalytical Chemistry* 408 (8): 2171–83. https://doi.org/10.1007/s00216-016-9321-6.

Müller, Simon. 2019. “Flexible Heuristic Algorithm for Automatic Molecule Fragmentation: Application to the UNIFAC Group Contribution Model.” *Journal of Cheminformatics* 11 (1): 57. https://doi.org/10.1186/s13321-019-0382-3.

O’Callaghan, Sean, David P. De Souza, Andrew Isaac, Qiao Wang, Luke Hodkinson, Moshe Olshansky, Tim Erwin, et al. 2012. “PyMS: A Python Toolkit for Processing of Gas Chromatography-Mass Spectrometry (GC-MS) Data. Application and Comparative Study of Selected Tools.” *BMC Bioinformatics* 13 (1): 115. https://doi.org/10.1186/1471-2105-13-115.

Olcese, Roberto, Vincent Carré, Frédéric Aubriet, and Anthony Dufour. 2013. “Selectivity of Bio-Oils Catalytic Hydrotreatment Assessed by Petroleomic and GC*GC/MS-FID Analysis.” *Energy & Fuels* 27 (4): 2135–45. https://doi.org/10.1021/ef302145g.

Panisko, Ellen, Thomas Wietsma, Teresa Lemmon, Karl Albrecht, and Daniel Howe. 2015. “Characterization of the Aqueous Fractions from Hydrotreatment and Hydrothermal Liquefaction of Lignocellulosic Feedstocks.” *Biomass and Bioenergy* 74 (March): 162–71. https://doi.org/10.1016/j.biombioe.2015.01.011.

Patwardhan, Pushkaraj R., Robert C. Brown, and Brent H. Shanks. 2011. “Understanding the Fast Pyrolysis of Lignin.” *ChemSusChem* 4 (11): 1629–36. https://doi.org/10.1002/cssc.201100133.

Sharma, Kamaldeep, Thomas Helmer Pedersen, Saqib Sohail Toor, Yves Schuurman, and Lasse Aistrup Rosendahl. 2020. “Detailed Investigation of Compatibility of Hydrothermal Liquefaction Derived Biocrude Oil with Fossil Fuel for Corefining to Drop-in Biofuels Through Structural and Compositional Analysis.” *ACS Sustainable Chemistry & Engineering* 8 (22): 8111–23. https://doi.org/10.1021/acssuschemeng.9b06253.

Sudibyo, Hanifrahmawan, Matteo Pecchi, and Jefferson William Tester. 2022. “Experimental-Based Mechanistic Study and Optimization of Hydrothermal Liquefaction of Anaerobic Digestates.” *Sustainable Energy & Fuels* 6 (9): 2314–29. https://doi.org/10.1039/D2SE00206J.

Sugumaran, Vatsala, Shanti Prakash, Emmandi Ramu, Ajay Kumar Arora, Veena Bansal, Vivekanand Kagdiyal, and Deepak Saxena. 2017. “Detailed Characterization of Bio-Oil from Pyrolysis of Non-Edible Seed-Cakes by Fourier Transform Infrared Spectroscopy (FTIR) and Gas Chromatography

Mass Spectrometry (GC–MS) Techniques." *Journal of Chromatography B* 1058 (July): 47–56. https://doi.org/10.1016/j.jchromb.2017.05.014.

Swain, Matt. 2017. "PubChemPy." Python. https://pubchempy.readthedocs.io/en/latest/guide/introduction.html.

Villadsen, Søren Ryom, Line Dithmer, Rasmus Forsberg, Jacob Becker, Andreas Rudolf, Steen Brummerstedt Iversen, Bo Brummerstedt Iversen, and Marianne Glasius. 2012. "Development and Application of Chemical Analysis Methods for Investigation of Bio-Oils and Aqueous Phase from Hydrothermal Liquefaction of Biomass." *Energy & Fuels* 26 (11): 6988–98. https://doi.org/10.1021/ef300954e.

Wang, Yinghao, Yehua Han, Wenya Hu, Dali Fu, and Gang Wang. 2020. "Analytical Strategies for Chemical Characterization of Bio-Oil." *Journal of Separation Science* 43 (1): 360–71. https://doi.org/10.1002/jssc.201901014.

Weininger, David. 1988. "SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules." *Journal of Chemical Information and Computer Sciences* 28 (1): 31–36. https://doi.org/10.1021/ci00057a005.

Yang, Jie, Quan (Sophia) He, Kenneth Corscadden, and Haibo Niu. 2018. "The Impact of Downstream Processing Methods on the Yield and Physiochemical Properties of Hydrothermal Liquefaction Bio-Oil." *Fuel Processing Technology* 178 (September): 353–61. https://doi.org/10.1016/j.fuproc.2018.07.006.

Zhu, Zhe, Xiangyu Guo, Lasse Rosendahl, Saqib Sohail Toor, Shuo Zhang, Zhiqiang Sun, Sensen Lu, Junying Zhao, Jinjun Yang, and Guanyi Chen. 2022. "Fast Hydrothermal Liquefaction of Barley Straw: Reaction Products and Pathways." *Biomass and Bioenergy* 165 (October): 106587. https://doi.org/10.1016/j.biombioe.2022.106587.