

2-7-2018

Automating Data Analysis for Two-Dimensional Gas Chromatography/Time-of-Flight Mass Spectrometry Non-Targeted Analysis of Comparative Samples

Ivan A. Titaley
Oregon State University


O. Maduka Ogba
Chapman University, ogba@chapman.edu

Leah Chibwe
Oregon State University

Eunha Hoh
San Diego State University

Paul H.-Y. Cheong
Oregon State University

Follow this and additional works at: https://digitalcommons.chapman.edu/sees_articles

 [next page for additional authors](#)
Part of the [Environmental Chemistry Commons](#), [Numerical Analysis and Scientific Computing Commons](#), [Other Chemistry Commons](#), [Other Computer Sciences Commons](#), [Other Life Sciences Commons](#), and the [Soil Science Commons](#)

Recommended Citation

Titaley IA, Ogba OM, Chibwe L, Hoh E, Cheong PHY, Simonich SLM. Automating data analysis for two-dimensional gas chromatography/time-of-flight mass spectrometry nontargeted analysis of comparative samples. *J Chromatog A*. 2018;1541:57–62. doi: 10.1016/j.chroma.2018.02.016

This Article is brought to you for free and open access by the Science and Technology Faculty Articles and Research at Chapman University Digital Commons. It has been accepted for inclusion in Biology, Chemistry, and Environmental Sciences Faculty Articles and Research by an authorized administrator of Chapman University Digital Commons. For more information, please contact laughtin@chapman.edu.

Automating Data Analysis for Two-Dimensional Gas Chromatography/ Time-of-Flight Mass Spectrometry Non-Targeted Analysis of Comparative Samples

Comments

NOTICE: this is the author's version of a work that was accepted for publication in *Journal of Chromatography A*. Changes resulting from the publishing process, such as peer review, editing, corrections, structural formatting, and other quality control mechanisms may not be reflected in this document. Changes may have been made to this work since it was submitted for publication. A definitive version was subsequently published in *Journal of Chromatography A*, volume 1541, in 2018. DOI: [10.1016/j.chroma.2018.02.016](https://doi.org/10.1016/j.chroma.2018.02.016)

The Creative Commons license below applies only to this version of the article.

Creative Commons License



This work is licensed under a [Creative Commons Attribution-Noncommercial-No Derivative Works 4.0 License](https://creativecommons.org/licenses/by-nc-nd/4.0/).

Copyright

Elsevier

Authors

Ivan A. Titaley, O. Maduka Ogba, Leah Chibwe, Eunha Hoh, Paul H.-Y. Cheong, and Staci L. Massey Simonich



HHS Public Access

Author manuscript

J Chromatogr A. Author manuscript; available in PMC 2019 March 16.

Published in final edited form as:

J Chromatogr A. 2018 March 16; 1541: 57–62. doi:10.1016/j.chroma.2018.02.016.

Automating Data Analysis for Two-Dimensional Gas Chromatography/Time-of-Flight Mass Spectrometry Non-Targeted Analysis of Comparative Samples

Ivan A. Titaley^{a,1,#}, O. Maduka Ogba^{a,b,#}, Leah Chibwe^{a,2}, Eunha Hoh^c, Paul H.-Y. Cheong^{a,*}, and Staci L. Massey Simonich^{a,d,*}

^aDepartment of Chemistry, Oregon State University, Corvallis, OR 97331, USA

^bDepartment of Chemistry, Pomona College, Claremont, CA, 91711 USA

^cGraduate School of Public Health, San Diego State University, San Diego, CA, 92182 USA

^dDepartment of Environmental and Molecular Toxicology, Oregon State University, Corvallis, OR, 97331 USA

Abstract

Non-targeted analysis of environmental samples, using comprehensive two-dimensional gas chromatography coupled with time-of-flight mass spectrometry (GC×GC/ToF-MS), poses significant data analysis challenges due to the large number of possible analytes. Non-targeted data analysis of complex mixtures is prone to human bias and is laborious, particularly for comparative environmental samples such as contaminated soil pre- and post-bioremediation. To address this research bottleneck, we developed *OCTpy*, a Python™ script that acts as a data reduction filter to automate GC×GC/ToF-MS data analysis from LECO® ChromaTOF® software and facilitates selection of analytes of interest based on peak area comparison between comparative samples. We used data from polycyclic aromatic hydrocarbon (PAH) contaminated soil, pre- and post-bioremediation, to assess the effectiveness of *OCTpy* in facilitating the selection of analytes that have *formed* or *degraded* following treatment. Using datasets from the soil extracts pre- and post-bioremediation, *OCTpy* selected, on average, 18% of the initial suggested analytes generated by the LECO® ChromaTOF® software Statistical Compare feature. Based on this list, 63–100% of the candidate analytes identified by a highly trained individual were also selected by *OCTpy*. This process was accomplished in several minutes per sample, whereas manual data analysis took several hours per sample. *OCTpy* automates the analysis of complex mixtures of comparative samples, reduces the potential for human error during heavy data handling and decreases data analysis time by at least tenfold.

*Corresponding Authors: paulc@science.oregonstate.edu, staci.simonich@oregonstate.edu phones: +1 (541) 737-6760 (PH-YC), +1 (541) 737-9194 (SLMS); fax: +1 (541) 737-0497 (SLMS).

¹Present address: Man-Technology-Environment Research Centre (MTM), School of Science and Technology, Örebro University, SE-70182 Örebro, Sweden

²Present address: Canada Centre for Inland Waters, Burlington, Ontario L7S 1A1, Canada

#IAT and OMO are co-first authors

Keywords

GC×GC/ToF-MS; Python™; Non-targeted analysis; LECO®; ChromaTOF®; Statistical Compare

1. Introduction

Comprehensive two-dimensional gas chromatography coupled with time-of-flight mass spectrometry (GC×GC/ToF-MS) is a versatile tool capable of elucidating organic analytes in complex environmental mixtures.[1–3] GC×GC/ToF-MS is also useful in non-targeted analysis[4–6] to help narrow down or identify analytes of interest, due to its higher peak capacity, sensitivity, and selectivity.[7–11] However, GC×GC/ToF-MS generates a large amount of data, which can be a major bottleneck in data analysis. Handling large volumes of data is not only time-consuming, particularly in non-targeted analysis, but is also prone to inter-laboratory discrepancies due to human error, regardless of the sample matrix. [3,7,12,13] For these reasons, an automated approach for GC×GC/ToF-MS data analysis is desired.

Historically, Microsoft® VBScripts,[1] C++,[2] MathWorks® MATLAB[14], and Guineu[15] have been the programming language or computer programs of choice for the development of data analysis scripts. Python™ has not been applied for the analysis of GC×GC/ToF-MS data despite its widespread use,[16–18] including for GC/MS data analysis.[19–21] Furthermore, most of the programming tools available for non-targeted analysis focus on environmental monitoring and not comparative samples, such as samples with and without treatment or exposure. An automated data analysis tool, which compares analyte concentrations before and after treatment or exposure to determine analyte *formation* or *degradation*, is also not currently available. Therefore, there is a need to automate GC×GC/ToF-MS data analysis of comparative samples.

In non-targeted analysis, a workflow to prioritize the identification and confirmation of candidate compounds is needed. Studies using GC×GC/ToF-MS analysis to screen for halogenated organic compounds in marine environments relied on halogenated isotopic clusters, fragmentation patterns indicating loss of halogen, and mass spectral library matches to identify, prioritize, and confirm candidate compounds [10,22,23]. The prioritization of chemicals can be based on several factors, including whether there is fragmentation data for known standards, electron impact (EI) mass spectral databases (such as NIST EI mass spectral library), previously published mass spectra data, and retention time match with known standards [11,23–25]. For EI, there are well developed strategies to predict compounds based on mass fragments, several fragments indicating unique functional groups, or chemical structures [26].

The objective of this study was to develop an automated method to streamline data analysis for GC×GC/ToF-MS non-targeted analysis of comparative samples by reducing the data analysis time. We developed a user-friendly Python™ script (*OCTpy*) that automates GC×GC/ToF-MS data interpretation by refining the list of suggested analytes generated by the LECO® ChromaTOF® software. In a previous study, coal tar-contaminated soil samples were aerobically remediated in a laboratory-scale bioreactor and measured for toxicity pre-

and post-bioremediation [27]. Although concentrations of parent-PAHs, which are environmental contaminants with human health implications,[28–30] *decreased* post-bioremediation, the soil toxicity *increased*, as evidenced by increase in genotoxicity from the *DT40* chicken lymphocyte bioassay and observed morphological malformations from the zebrafish (*Danio rerio*) developmental toxicity testing.[27] We used *OCTpy* to determine analytes *formed* during bioremediation of soil samples with data generated by GC×GC/ToF-MS.[27] We benchmarked the *OCTpy* results with those obtained through human manual analysis to establish the reliability of *OCTpy* results and the effectiveness of *OCTpy* in reducing GC×GC/ToF-MS data analysis time.

2. Material and Methods

2.1 Sample Extraction and Analysis

It was hypothesized that the increase in soil toxicity was due to the formation of oxygenated transformation products of 3- and 4-ringed parent PAHs and other unknown analytes during bioremediation.[27] To test this hypothesis, extracts of pre- and post-bioremediation soil samples were analyzed using GC×GC/ToF-MS. Briefly, soil samples were extracted in hexane:acetone (75:25 v/v) using pressurized liquid extraction and fractionated into six fractions (A–F) based on polarity.[27] Laboratory blanks (consisting of sodium sulfate) also went through the same extraction procedures. To analyze analytes of interest that were not GC amenable, such as hydroxy-PAHs, aliquots of soil fractions were derivatized with *N-tert*-Butyldimethylsilyl-*N*-methyltrifluoroacetamide (MTBSTFA).[31] Both underivatized and derivatized fractions were analyzed in triplicate. Further details of the analytical procedures are available elsewhere.[11,27]

Pre- and post-bioremediation soil fraction extracts, along with blanks, were analyzed using an Agilent 7890 GC (Palo Alto, CA) coupled to a LECO® Pegasus® 4D time-of-flight mass spectrometer (St. Joseph, MI) with Restek Rtx-5 (35 m × 0.25 mm I.D. × 0.1 μm film thickness with 5-m integrated guard column) and Rxi-17 (1 m × 0.1 mm I.D. × 0.1 μm film thickness) (Bellefonte, PA) as the first and second dimension columns, respectively. [11,32,33] The injection volume for each sample was 2 μL, and the inlet temperature was held at 300 °C. For the first dimension, the oven was held at 60 °C for 1 min, ramped at 6 °C/min to 300 °C, held for 3 min, ramped at 20 °C/min to a final temperature of 320 °C, and held for 15 min. For the second dimension, the oven was first held at 85 °C for 1 min, ramped at 6 °C/min to 320 °C, held for 3 min, ramped at 20 °C/min to 340 °C, and held for 15 min. The non-moving quad-jet dual-stage modulator temperature was 35 °C higher than the temperature in the first-dimension column, with a modulation period of 3.5 s. Transfer line and ion source were kept at 285 °C and 250 °C, respectively.

GC×GC/ToF-MS data for the soil bioremediation extracts were analyzed using LECO® ChromaTOF® software (St. Joseph, MI)[7,14,34] version 4.50. LECO® ChromaTOF® software has peak alignment, baseline correction, and peak deconvolution capability, but the program does not automate peak area difference calculations in chromatograms from comparative samples, resulting in additional data analysis time. Matches between the second dimension peaks and peaks from individual replicates were set at 500 (50%).[35] The minimum signal-to-noise (S/N) ratio for peak finding was 50. The LECO® ChromaTOF®

software's add-in feature, Statistical Compare (SC) (St. Joseph, MI), [34,36,37] was used in the soil bioremediation study. The LECO[®] ChromaTOF[®] software returned a list of suggested analytes based on mass spectral similarity matches to the National Institute of Standard and Technology (NIST) 2011 library. Based on the SC's feature that identifies statistical differences in groups, we compared peaks assigned to three groups: blank, pre-, and post-bioremediation soil extracts. The output files, containing potential analytes of interest, were exported as text files (*.txt). Henceforth in this manuscript, "suggested analytes" refers to the list of analytes from the LECO[®] ChromaTOF[®] software's SC output based on MS NIST 2011 library matches, while "candidate analytes" refers to the list of analytes from manual analysis or as an *OCTpy* output.

3. Theory

3.1 Manual Data Analysis

Figure 1 shows the workflow for manual data analysis. If a peak in the post-bioremediation extract increased by 1.5 fold, the analyte was interpreted as a candidate analyte formed during bioremediation and could potentially be responsible for the observed increase in toxicity post-bioremediation. These peaks were visually inspected, and only those peaks that were chromatographically resolved, and without severe tailing, were selected as candidate peaks. This manual workflow was repeated for each analyte, in each derivatized and underivatized fraction, and took many hours to accomplish.

3.2 Automated Data Analysis

Figure 2 shows the workflow for automated data analysis with *OCTpy*, which relies on SC and acquired data input files from LECO[®] ChromaTOF[®] software. The SC input files contain data on the average peak area, for every suggested analyte, in every treatment group. For the purpose of this script, it is necessary to export the following four properties into a text file for every sample: (1) Analyte name ("*Name*"), (2) Quantifying mass-to-charge (*m/z*) ion ("*Mass*"), (3) Average quantifying peak area ("*Area Average*"), and (4) Average retention time ("*R. T. Average (s)*"). Having the area and RT average parameters in the datasets are advantageous because different experimental designs can have different numbers of replicates. There were three SC input files for every fraction in the soil bioremediation study (*i.e.*, blank, pre-, and post-bioremediation files).

OCTpy also requires an acquired sample input file prior to the SC feature analysis from one replicate, in one of the treatment groups, in order for the script to run. Serving as a quality control step, the acquired file is a reference file that verifies the presence and precision of the quantifying *m/z* ion and second dimension RT for each suggested analyte. The second dimension RT is used because there can be variations in the first dimension RT that arises from modulation time differences if different chromatogram slice is used as the base. The following three properties are pertinent for the script: (1) Molecular weight of the suggested analyte ("*Exact Mass*"), (2) First and second dimension retention times ("*R. T. (s)*"), and (3) Quantifying MS ion ("*Quant Masses*"). We used a replicate from the post-bioremediation sample for the acquired sample input file. Instructions for obtaining the SC input and the acquired sample input files for *OCTpy* are available in Appendix A.

4. Results and Discussions

OCTpy automates peak area comparisons of suggested analytes between comparative samples from the SC feature in LECO® ChromaTOF® software, thereby facilitating analyte selection by the analyst and shortening the data analysis times. Depending on the number of candidate analytes identified by the SC feature, the script returned results within seconds (Intel® Core™2 Quad CPU, 8 GB RAM, Windows® 7, 64-bit), whereas manual data analysis required several hours for the same candidate analytes. Although preparing input files for *OCTpy* can take several minutes, analysts would still spend less time analyzing the peak area data by using *OCTpy*, compared to manual analysis. This is because *OCTpy* filters out suggested analytes that do not satisfy the desired criteria, such as those for which peak areas decreased following soil bioremediation. Following peak area comparison in a manual analysis, the analyst must meet additional criteria for identification by comparing the peak shapes of candidate analytes (Figure 1). However, we found that peak shape did not alter the results of *OCTpy* outputs and peak shape is not correlated with the toxicity of an analyte. Therefore, peak shape was excluded from the *OCTpy* script. The choice to either further review the peak shapes of candidate analytes from *OCTpy* with LECO® ChromaTOF®, or to use the output from *OCTpy* as the final result, is left to the discretion of the analyst. Once peak areas are compared, *OCTpy* parses through the acquired sample input file and checks each candidate analyte for the presence of the quantifying m/z ion or second dimension RT to the nearest tenth of a second (with an adjusted range of ± 0.1 s to accommodate RT shift). For instance, if a candidate analyte has RTs of 100.789 s and 2.34 s, and the reference dataset contains RTs of 100.789 s and 2.429 s, then the candidate analyte will be included in the output. We successfully tested *OCTpy* on computers with Microsoft Windows®, Mac®, and Linux® operating systems. The *OCTpy* source code and instructions for running *OCTpy* from the Windows® command prompt are available in Supplementary materials section.

OCTpy outputs are available in the SI. We contrasted the candidate analytes identified using *OCTpy* to those identified by manual analysis, which included peak area comparison and individual peak review. Only analytes with chemical names (*i.e.*, minimum similarity match with 2011 NIST library before name assigned > 700 (70%)) were selected for further analysis. There were a total of eight paired datasets, each consisting of a blank, pre-, and post-bioremediation samples from four soil extract fractions. Only four soil extract fractions were analyzed using GC×GC/ToF-MS, because increases in toxicity post-bioremediation were only measured in these four fractions.[11,27] Four of the eight datasets were from derivatized extracts, while four were from underivatized extracts. Among the eight paired datasets, the number of suggested analytes from the SC feature ranged from 170 (Fraction D) to 1,650 (Fraction C Derivatized) (Table 1). From the initial SC lists, *OCTpy* selected 20 (Fraction D) to 307 (Fraction D Derivatized) analytes having increased peak areas following bioremediation. As mentioned, once a list of candidate analytes is generated by *OCTpy*, the analyst is left with the option to either further examine the peak shape of each suggested analyte with the LECO® ChromaTOF® software or to proceed with the existing data and to look for specific groups of analytes, such as PAH transformation products.

In contrast, manual analysis identified 6 (Fraction F) to 27 (Fraction D Derivatized) candidate analytes with increased peak areas following bioremediation. The manual lists of

candidate analytes were shorter than the *OCTpy* list of candidate analytes because they underwent manual peak shape review with the LECO® ChromaTOF® software and *OCTpy* did not. For all eight data sets, *OCTpy* results contained 63–100% of the final list of candidate analytes ($n = 87$) generated by manual data analysis (Table 1). *OCTpy* generated more candidate analytes than manual data analysis of the initial list of suggested analytes from the SC feature, because peak shape was not included in the analyte selection criteria by *OCTpy*. *OCTpy* was able to select, on average, 18% of candidate analytes from the initial SC list output in a matter of several minutes. Even if an *OCTpy* user chooses to examine the peak shapes of the candidate analytes, the number of analytes to be examined is significantly reduced by executing *OCTpy* first, which results in decreased data analysis time. For example, the number of suggested analytes, from the SC feature analysis, in fraction D-Derivatized were reduced four fold once *OCTpy* was executed, which allowed the analyst to start the analysis with much fewer suggested analytes.

Twenty nine candidate analytes from manual analysis of the eight soil bioremediation data sets were not found in the *OCTpy* outputs. Several factors contributed to this. First, no peak areas (no value) in the chromatograms were replaced with the value “0” manually, whereas in the LECO® ChromaTOF® software SC input files, used in the *OCTpy* script, the fields were left blank. For example, for a dataset in which a given treatment group has three peaks with area values and one peak without, the analyst would manually calculate the average with $n = 4$, whereas LECO® ChromaTOF® and, consequently, *OCTpy*, would calculate the average with $n = 3$. This difference in how the average peak area was calculated explained why the seven candidate analytes that were selected manually in Fraction C were omitted by *OCTpy*. Given that different analysts treat missing values differently, it is imperative to create a standard method to analyze data that contain peak areas with no values. We recommend to using empty fields instead of inserting “0”. Since LECO® ChromaTOF® software calculates average peak areas without replacing any with zero, we propose *OCTpy* as the method of choice when analyzing LECO® ChromaTOF® software data, because it standardizes the calculation of average peak area.

Following close examination of the raw data, *OCTpy* excluded the remaining twenty two candidate analytes identified by the manual data analysis for one of the two following reasons (both of which were largely attributed to human error during data analysis): 1) the peak areas in the blank fractions were higher than those in the pre-bioremediation fractions, or 2) the peak areas in the pre-bioremediation fractions were higher than the post-bioremediation fractions (Appendix B). Detection of these discrepancies demonstrated *OCTpy*'s ability to account for human errors prior to qualitative identification of candidate analyte structures. However, future studies that use *OCTpy* should consider an assessment where manual and automated methods are compared, using a number of authentic standards with two different known concentrations, to reduce the potential for discrepancy between the two methods.

OCTpy's capacity to reduce the number of candidate analytes within several minutes, along with its ability to reduce the potential for human error, represents a significant improvement over manual analysis that can take several hours. *OCTpy* gives the user the flexibility to not only identify analytes with increased peak areas following treatment in comparative

samples, but also those with decreased peak areas. In addition, the user can also choose to either analyze differences between two groups, such as pre- and post-bioremediation, or three groups, such as blank, pre-, and post-bioremediation.

Once candidate compounds are identified, the analyst must confirm the structure of compounds using pure standards [24,38,39]. Using standards, Chibwe *et al.* identified N-(5-amino-4-cyano-1-pyrazolyl)phthalimide and 4-methylphthalic anhydride as the two compounds in fraction F with increased peak area post-bioremediation based on GC×GC/ToF-MS analysis.

5. Conclusions

OCTpy is a python script designed to automate analysis of GC×GC/ToF-MS data obtained from LECO® ChromaTOF® software. *OCTpy* uses SC feature input files that were generated by the LECO® ChromaTOF® software. Automated data analysis can now be completed within several minutes using *OCTpy*, making *OCTpy* an efficient platform that can be adopted by analysts who are using GC×GC/ToF-MS and the SC feature from the LECO® ChromaTOF® software. Because the list of candidate analytes are curated using the SC feature, only the most statistically significant analytes are included. However, the analyst should ensure that the MS fragmentation pattern of the candidate analyte is the same as the NIST EI mass spectral library.

While the decision to review peak shape following *OCTpy* analysis is left to the analysts, *OCTpy* provides assistance by making the analysis of large datasets practical and less time-consuming [41]. In cases when there are a significant number of candidate compounds that need to be confirmed for their toxicity, high throughput screening, such as the zebrafish developmental toxicity testing [42,43], or *in silico* search using the U.S. Environmental Protection Agency's Tox 21 approach [44,45], are useful to reduce testing time. *OCTpy* also reduces human error, which is considered the most common source of error in quantitative analysis.[46] *OCTpy*'s capabilities are also unique, because the script aids researchers in data analysis and screening of candidate analytes with *increased* or *decreased* peak areas for comparative samples. In the spirit of open-source,[47] *OCTpy* is available on request, allowing further developments by users who are interested in examining the source code. Integrating *OCTpy* into the overall workflow of GC×GC/ToF-MS for non-targeted analysis[24] can also provide a standard method, across laboratories and research groups, for data analysis.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by the National Institute of Environmental Health Science [grant numbers P30ES00210, P01ES021921, P42ES016465, T32ES007060]; National Science Foundation [grant number AGS-1411214]; OSU Dorothy Ramon Barnes Fellowship; Pomona College Robbins Post-doctoral Fellowship; and OSU Stone Family.

PH-YC is the Bert and Emelyn Christensen Professor of Chemistry and OMO is the Robbins Postdoctoral Fellow in Chemistry. The authors also thank Dr. Cleo Davie-Martin and Dr. Mary Leonard for reviewing the manuscript and providing valuable inputs to the draft.

References

1. Hilton DC, Jones RS, Sjödin A. A method for rapid, non-targeted screening for environmental contaminants in household dust. *J Chromatogr A*. 2010; 1217:6851–6856. DOI: 10.1016/j.chroma.2010.08.039 [PubMed: 20864112]
2. Kallio M, Jussila M, Rissanen T, Anttila P, Hartonen K, Reissell A, Vreuls R, Adahchour M, Hyötyläinen T. Comprehensive two-dimensional gas chromatography coupled to time-of-flight mass spectrometry in the identification of organic compounds in atmospheric aerosols from coniferous forest. *J Chromatogr A*. 2006; 1125:234–243. DOI: 10.1016/j.chroma.2006.05.050 [PubMed: 16782114]
3. Parastar H, Radovi JR, Jalali-Heravi M, Diez S, Bayona JM, Tauler R. Resolution and Quantification of Complex Mixtures of Polycyclic Aromatic Hydrocarbons in Heavy Fuel Oil Sample by Means of GC × GC-TOFMS Combined to Multivariate Curve Resolution. *Anal Chem*. 2011; 83:9289–9297. DOI: 10.1021/ac201799r [PubMed: 22077766]
4. Brack W, Ait-Aissa S, Burgess RM, Busch W, Creusot N, Di Paolo C, Escher BI, Mark Hewitt L, Hilscherova K, Hollender J, Hollert H, Jonker W, Kool J, Lamoree M, Muschket M, Neumann S, Rostkowski P, Ruttkies C, Schollee J, Schymanski EL, Schulze T, Seiler T-B, Tindall AJ, De Aragão Umbuzeiro G, Vrana B, Krauss M. Effect-Directed Analysis Supporting Monitoring of Aquatic Environments — An In-Depth Overview. *Sci Total Environ*. 2016; 544:1073–1118. DOI: 10.1016/j.scitotenv.2015.11.102 [PubMed: 26779957]
5. Reichenbach SE, Tian X, Cordero C, Tao Q. Features for non-targeted cross-sample analysis with comprehensive two-dimensional chromatography. *J Chromatogr A*. 2012; 1226:140–148. DOI: 10.1016/j.chroma.2011.07.046 [PubMed: 21855071]
6. Hoh E, Lehotay SJ, Mastovska K, Ngo HL, Vetter W, Pangallo KC, Reddy CM. Capabilities of Direct Sample Introduction-Comprehensive Two-Dimensional Gas Chromatography-Time-of-Flight Mass Spectrometry to Analyze Organic Chemicals of Interest in Fish Oils. *Environ Sci Technol*. 2009; 43:3240–3247. DOI: 10.1021/es803486x [PubMed: 19534141]
7. Almstetter MF, Appel IJ, Gruber MA, Lottaz C, Timischl B, Spang R, Dettmer K, Oefner PJ. Integrative Normalization and Comparative Analysis for Metabolic Fingerprinting by Comprehensive Two-Dimensional Gas Chromatography-Time-of-Flight Mass Spectrometry. *Anal Chem*. 2009; 81:5731–5739. DOI: 10.1021/ac900528b [PubMed: 19522528]
8. Murray JA. Qualitative and quantitative approaches in comprehensive two-dimensional gas chromatography. *J Chromatogr A*. 2012; 1261:58–68. DOI: 10.1016/j.chroma.2012.05.012 [PubMed: 22647189]
9. Kim S, Koo I, Fang A, Zhang X. Smith-Waterman peak alignment for comprehensive two-dimensional gas chromatography-mass spectrometry. *BMC Bioinformatics*. 2011; 12:235. doi: 10.1186/1471-2105-12-235 [PubMed: 21676240]
10. Shaul NJ, Dodder NG, Aluwihare LI, Mackintosh SA, Maruya KA, Chivers SJ, Danil K, Weller DW, Hoh E. Nontargeted Biomonitoring of Halogenated Organic Compounds in Two Ecotypes of Bottlenose Dolphins (*Tursiops truncatus*) from the Southern California Bight. *Environ Sci Technol*. 2015; 49:1328–1338. DOI: 10.1021/es505156q [PubMed: 25526519]
11. Chibwe L, Davie-Martin CL, Aitken MD, Hoh E, Massey Simonich SL. Identification of Polar Transformation Products and High Molecular Weight Polycyclic Aromatic Hydrocarbons (PAHs) in Contaminated Soil following Bioremediation. *Sci Total Environ*. 2017; 599–600:1099–1107. DOI: 10.1016/j.scitotenv.2017.04.190
12. Jeong J, Shi X, Zhang X, Kim S, Shen C. Model-based peak alignment of metabolomic profiling from comprehensive two-dimensional gas chromatography mass spectrometry. *BMC Bioinformatics*. 2012; 13:27. doi: 10.1186/1471-2105-13-27 [PubMed: 22316124]
13. Seeley JV, Seeley SK. Multidimensional Gas Chromatography: Fundamental Advances and New Applications. *Anal Chem*. 2013; 85:557–578. DOI: 10.1021/ac303195u [PubMed: 23137217]

14. Oh C, Huang X, Regnier FE, Buck C, Zhang X. Comprehensive two-dimensional gas chromatography/time-of-flight mass spectrometry peak sorting algorithm. *J Chromatogr A*. 2008; 1179:205–215. DOI: 10.1016/j.chroma.2007.11.101 [PubMed: 18093607]
15. Castillo S, Mattila I, Miettinen J, Orešič M, Hyötyläinen T. Data Analysis Tool for Comprehensive Two-Dimensional Gas Chromatography/Time-of-Flight Mass Spectrometry. *Anal Chem*. 2011; 83:3058–3067. DOI: 10.1021/ac103308x [PubMed: 21434611]
16. Pasi M, Tiberti M, Arrigoni A, Papaleo E. xPyder: A PyMOL Plugin To Analyze Coupled Residues and Their Networks in Protein Structures. *J Chem Inf Model*. 2012; 52:1865–1874. DOI: 10.1021/ci300213c [PubMed: 22721491]
17. Cao DS, Liang YZ, Yan J, Tan GS, Xu QS, Liu S. PyDPI: Freely Available Python Package for Chemoinformatics, Bioinformatics, and Chemogenomics Studies. *J Chem Inf Model*. 2013; 53:3086–3096. DOI: 10.1021/ci400127q [PubMed: 24047419]
18. Li P, Merz KM. MCPB.py: A Python Based Metal Center Parameter Builder. *J Chem Inf Model*. 2016; 56:599–604. DOI: 10.1021/acs.jcim.5b00674 [PubMed: 26913476]
19. O’Callaghan S, De Souza DP, Isaac A, Wang Q, Hodkinson L, Olshansky M, Erwin T, Appelbe B, Tull DL, Roessner U, Bacic A, McConville MJ, Likiš VA. PyMS: a Python toolkit for processing of gas chromatography-mass spectrometry (GC-MS) data. Application and comparative study of selected tools. *BMC Bioinformatics*. 2012; 13:115.doi: 10.1186/1471-2105-13-115 [PubMed: 22647087]
20. Bald T, Barth J, Niehues A, Specht M, Hippler M, Fufezan C. pymzML—Python module for high-throughput bioinformatics on mass spectrometry data. *Bioinformatics*. 2012; 28:1052–1053. DOI: 10.1093/bioinformatics/bts066 [PubMed: 22302572]
21. Strohalm M, Kavan D, Novák P, Volný M, Havlíček V. mMass 3: A Cross-Platform Software Environment for Precise Analysis of Mass Spectrometric Data. *Anal Chem*. 2010; 82:4648–4651. DOI: 10.1021/ac100818g [PubMed: 20465224]
22. Hoh E, Dodder NG, Lehotay SJ, Pangallo KC, Reddy CM, Maruya KA. Nontargeted Comprehensive Two-Dimensional Gas Chromatography/Time-of-Flight Mass Spectrometry Method and Software for Inventorying Persistent and Bioaccumulative Contaminants in Marine Environments. *Environ Sci Technol*. 2012; 46:8001–8008. DOI: 10.1021/es301139q [PubMed: 22712571]
23. Alonso MB, Maruya KA, Dodder NG, Lailson-Brito J, Azevedo A, Santos-Neto E, Torres JPM, Malm O, Hoh E. Nontargeted Screening of Halogenated Organic Compounds in Bottlenose Dolphins (*Tursiops truncatus*) from Rio de Janeiro, Brazil. *Environ Sci Technol*. 2017; 51:1176–1185. DOI: 10.1021/acs.est.6b04186 [PubMed: 28055195]
24. Chibwe L, Titaley IA, Hoh E, Simonich SLM. Integrated Framework for Identifying Toxic Transformation Products in Complex Environmental Mixtures. *Environ Sci Technol Lett*. 2017; 4:32–43. DOI: 10.1021/acs.estlett.6b00455
25. López SH, Ulaszewska MM, Hernando MD, Bueno MJM, Gómez MJ, Fernández-Alba AR. Post-Acquisition Data Processing for the Screening of Transformation Products of Different Organic Contaminants. Two-year Monitoring of River Water using LC-ESI-QToF-MS and GCxGC-EI-ToF-MS. *Environ Sci Pollut Res*. 2014; 21:12583–12604. DOI: 10.1007/s11356-014-3187-y
26. McLafferty, FW., Tureček, F. Interpretation of Mass Spectra. 4. University Science Books; Sausalito, CA: 1993.
27. Chibwe L, Geier MC, Nakamura J, Tanguay RL, Aitken MD, Simonich SLM. Aerobic Bioremediation of PAH Contaminated Soil Results in Increased Genotoxicity and Developmental Toxicity. *Environ Sci Technol*. 2015; 49:13889–13898. DOI: 10.1021/acs.est.5b00499 [PubMed: 26200254]
28. Titaley IA, Chlebowski A, Truong L, Tanguay RL, Massey Simonich SL. Identification and Toxicological Evaluation of Unsubstituted PAHs and Novel PAH Derivatives in Pavement Sealcoat Products. *Environ Sci Technol Lett*. 2016; 3:234–242. DOI: 10.1021/acs.estlett.6b00116
29. Chlebowski AC, Garcia GR, Du L, JK, Bisson WH, Truong L, Simonich M, SL, Tanguay RL. Mechanistic Investigations Into the Developmental Toxicity of Nitrated and Heterocyclic PAHs. *Toxicol Sci*. 2017; 157:246–259. DOI: 10.1093/toxsci/kfx035 [PubMed: 28186253]

30. Wang W, Jariyasopit N, Schrlau J, Jia Y, Tao S, Yu TW, Dashwood RH, Zhang W, Wang X, Simonich SLM. Concentration and Photochemistry of PAHs, NPAHs, and OPAHs and Toxicity of PM_{2.5} during the Beijing Olympic Games. *Environ Sci Technol*. 2011; 45:6887–6895. DOI: 10.1021/es201443z [PubMed: 21766847]
31. Motorykin O, Santiago-Delgado L, Rohlman D, Schrlau JE, Harper B, Harris S, Harding A, Kile ML, Massey Simonich SL. Metabolism and excretion rates of parent and hydroxy-PAHs in urine collected after consumption of traditionally smoked salmon for Native American volunteers. *Sci Total Environ*. 2015; 514:170–177. DOI: 10.1016/j.scitotenv.2015.01.083 [PubMed: 25659315]
32. Manzano C, Hoh E, Simonich SLM. Improved Separation of Complex Polycyclic Aromatic Hydrocarbon Mixtures Using Novel Column Combinations in GC × GC/ToF-MS. *Environ Sci Technol*. 2012; 46:7677–7684. DOI: 10.1021/es301790h [PubMed: 22769970]
33. Manzano C, Hoh E, Simonich SLM. Quantification of complex polycyclic aromatic hydrocarbon mixtures in standard reference materials using comprehensive two-dimensional gas chromatography with time-of-flight mass spectrometry. *J Chromatogr A*. 2013; 1307:172–179. DOI: 10.1016/j.chroma.2013.07.093 [PubMed: 23932031]
34. Almstetter MF, Appel IJ, Dettmer K, Gruber MA, Oefner PJ. Comparison of two algorithmic data processing strategies for metabolic fingerprinting by comprehensive two-dimensional gas chromatography–time-of-flight mass spectrometry. *J Chromatogr A*. 2011; 1218:7031–7038. DOI: 10.1016/j.chroma.2011.08.006 [PubMed: 21871627]
35. Prebihalo S, Brockman A, Cochran J, Dorman FL. Determination of Emerging Contaminants in Wastewater Utilizing Comprehensive Two-Dimensional Gas-Chromatography Coupled with Time-of-Flight Mass Spectrometry. *J Chromatogr A*. 2015; 1419:109–115. DOI: 10.1016/j.chroma.2015.09.080 [PubMed: 26442816]
36. Parsons BA, Marney LC, Siegler WC, Hoggard JC, Wright BW, Synovec RE. Tile-Based Fisher Ratio Analysis of Comprehensive Two-Dimensional Gas Chromatography Time-of-Flight Mass Spectrometry (GC × GC–TOFMS) Data Using a Null Distribution Approach. *Anal Chem*. 2015; 87:3812–3819. DOI: 10.1021/ac504472s [PubMed: 25785933]
37. Stadler S, Stefanuto PH, Brokl M, Forbes SL, Focant JF. Characterization of Volatile Organic Compounds from Human Analogue Decomposition Using Thermal Desorption Coupled to Comprehensive Two-Dimensional Gas Chromatography–Time-of-Flight Mass Spectrometry. *Anal Chem*. 2013; 85:998–1005. DOI: 10.1021/ac302614y [PubMed: 23215054]
38. Brack W. Effect-Directed Analysis: A Promising Tool for the Identification of Organic Toxicants in Complex Mixtures? *Anal Bioanal Chem*. 2003; 377:397–407. DOI: 10.1007/s00216-003-2139-z [PubMed: 12904950]
39. Brack W, Schmitt-Jansen M, Machala M, Brix R, Barceló D, Schymanski E, Streck G, Schulze T. How to confirm identified toxicants in effect-directed analysis. *Anal Bioanal Chem*. 2008; 390:1959–1973. DOI: 10.1007/s00216-007-1808-8 [PubMed: 18224304]
41. Lehotay SJ, Koesukwiwat U, van der Kamp H, Mol HGJ, Leepipatpiboon N. Qualitative Aspects in the Analysis of Pesticide Residues in Fruits and Vegetables Using Fast, Low-Pressure Gas Chromatography–Time-of-Flight Mass Spectrometry. *J Agric Food Chem*. 2011; 59:7544–7556. DOI: 10.1021/jf104606j [PubMed: 21452898]
42. Truong L, Reif DM, Mary LS, Geier MC, Truong HD, Tanguay RL. Multidimensional in vivo hazard assessment using zebrafish. *Toxicol Sci*. 2014; 137:212–233. DOI: 10.1093/toxsci/kft235 [PubMed: 24136191]
43. Chlebowski AC, La Du JK, Truong L, Massey Simonich SL, Tanguay RL. Investigating the application of a nitroreductase-expressing transgenic zebrafish line for high-throughput toxicity testing. *Toxicol Rep*. 2017; 4:202–210. DOI: 10.1016/j.toxrep.2017.04.005 [PubMed: 28758069]
44. Wambaugh JF, Wang A, Dionisio KL, Frame A, Egeghy P, Judson R, Setzer RW. High Throughput Heuristics for Prioritizing Human Exposure to Environmental Chemicals. *Environ Sci Technol*. 2014; 48:12760–12767. DOI: 10.1021/es503583j [PubMed: 25343693]
45. Richard AM, Judson RS, Houck KA, Grulke CM, Volarath P, Thillainadarajah I, Yang C, Rathman J, Martin MT, Wambaugh JF, Knudsen TB, Kancherla J, Mansouri K, Patlewicz G, Williams AJ, Little SB, Crofton KM, Thomas RS. ToxCast Chemical Landscape: Paving the Road to 21st Century Toxicology. *Chem Res Toxicol*. 2016; 29:1225–1251. DOI: 10.1021/acs.chemrestox.6b00135 [PubMed: 27367298]

46. Lehotay SJ, Mastovska K, Amirav A, Fialkov AB, Martos PA, de Kok A, Fernández-Alba AR. Identification and confirmation of chemical residues in food by chromatography-mass spectrometry and other techniques. *TrAC Trends Anal Chem.* 2008; 27:1070–1090. DOI: 10.1016/j.trac.2008.10.004
47. Dryden MDM, Fobel R, Fobel C, Wheeler AR. Upon the Shoulders of Giants: Open-Source Hardware and Software in Analytical Chemistry. *Anal Chem.* 2017; 89:4330–4338. DOI: 10.1021/acs.analchem.7b00485 [PubMed: 28379683]

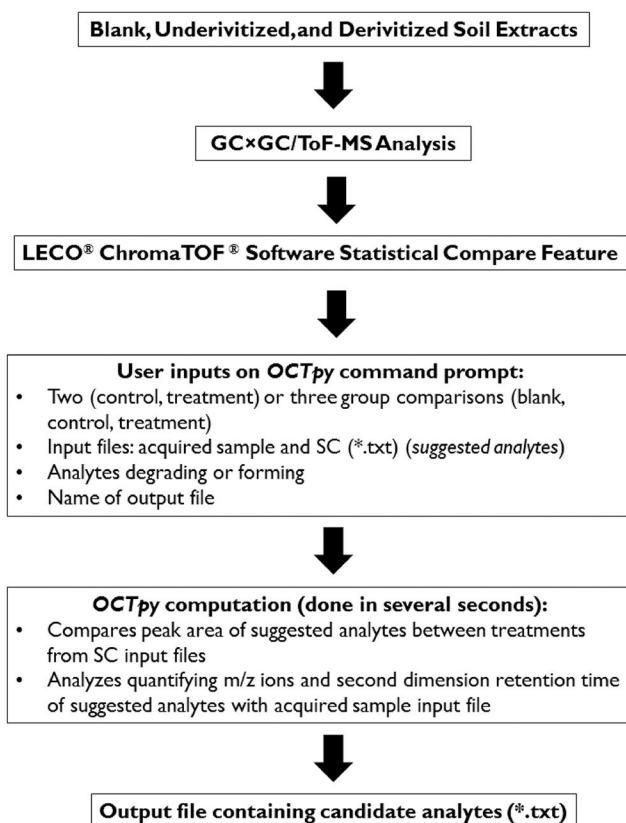


Figure 1. Manual Analysis Workflow

The manual workflow used to select candidate analytes from a given GCxGC/ToF-MS data output analysis from the LECO® ChromaTOF® software's Statistical Compare feature. This process can take several hours per sample, depending on the complexity of the sample.

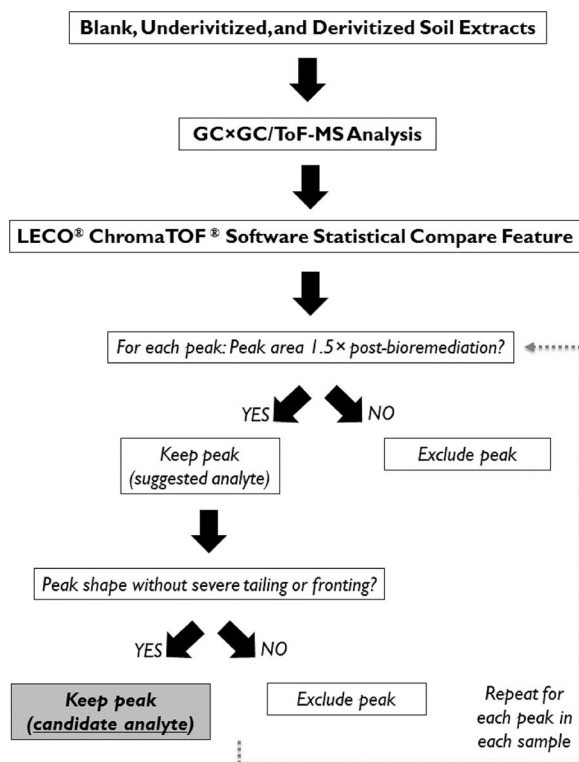


Figure 2. Automated Analysis Workflow

The automated workflow used to select candidate analytes from a given GCxGC/ToF-MS data output analysis from the LECO® ChromaTOF® software's Statistical Compare feature. This process only take several minutes per sample, depending on the complexity of the sample.

Table 1
Manuals vs. Automated Result Comparisons

Comparisons of candidate analytes selected by either manual analysis or *OCTpy* from the the LECO® ChromaTOF® software's Statistical Compare (SC) feature list. Percent match between manual analysis and *OCTpy* was calculated based on the percent of the ratio between the number of candidate analytes that were selected in both *OCTpy* and manual analysis. For example, in fraction D, although *OCTpy* selected 20 candidate analytes, the 7 candidate analytes selected from manual analysis were also included in the *OCTpy* candidate analytes list, which resulted in 100% match.

Remediated Soil Fractions	Initial Number of Suggested Analytes from SC	Number of Candidate Analytes Identified by Manual Analysis	Number of Candidate Analytes Identified by <i>OCTpy</i>	Percent Match between Manual Analysis and <i>OCTpy</i> (%)
C	205	23	38	70
C-Derivatized	1,650	24	277	63
D	170	7	20	100
D-Derivatized	1,240	27	307	63
E	1,060	10	195	90
E-Derivatized	1,010	12	168	83
F	750	6	93	83
F-Derivatized	560	7	151	100