JOURNAL OF CHEMICAL EDUCATION

pubs.acs.org/jchemeduc

Technology Report

Data Functionalization for Gas Chromatography in Python

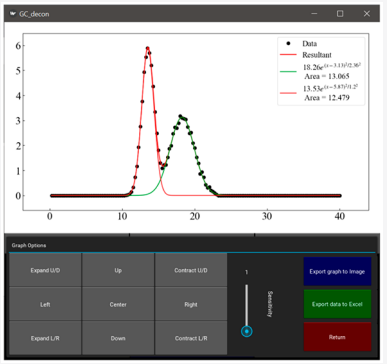
Michael Green and Xiaobo Chen*

Cite This: *J. Chem. Educ.* 2020, 97, 1172−1175

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: For undergraduate students to be prepared for graduate school and industry, it is imperative that they understand how to merge the theoretical insights gleaned through their undergraduate education with the raw data sets acquired through materials analysis. Thus, the ability to implement data analysis is a vital skill that students should develop. Furthermore, students should be fluent in methodologies that can translate to domains beyond their undergraduate curriculum. In this technology report, we demonstrate data functionalization in the Python programming language via data derived from gas chromatography. The programming approach to data analysis is designed to be flexible in order to allow students to take the lessons learned herein and apply them to novel systems outside of the experiment and outside of the academy.



KEYWORDS: *Upper-Division Undergraduate, Physical Chemistry, Interdisciplinary/Multidisciplinary, Computer-Based Learning, Gas Chromatography*

INTRODUCTION

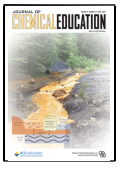
The ability of chemists to harness the full computational capacity of computers is becoming imperative in the current technological age. Therefore, equipping students to make use of data-driven methods provides them with the tools and understanding necessary to thrive in industry and postbaccalaureate studies in academia. In the current undergraduate curriculum, much effort is spent on instilling concepts and models of chemical systems. However, simply understanding concepts and models is oftentimes insufficient when chemists begin to work with raw, unadulterated data. In experimentation, data analysis and processing are often unavoidable. Thus, it is necessary that students of chemistry understand the tools they have at their disposal in order to assist in their comprehension and understanding of how to analyze experimental data sets.

Furthermore, the current curriculum needs to keep in mind the accessibility students have to computational development environments outside of the academy. Many times, university curricula utilize developing environments that are convenient to teach for niche topics, though such convenience comes at a price. Students pay the brunt of this cost, learning computational skills and techniques in developing environments to which they no longer have access upon leaving the academy. Such a system of learning is ultimately unaccommodating to students, who upon graduation are then required to either choose to purchase developing environments at a nontrivial cost or learn outside the academy an entirely new computational system from the ground up.

The purpose of this report is to mitigate against these two core issues. Specifically, many of the experiments in upper-division physical chemistry laboratory courses can be used to introduce students to data-driven methods via the Python programming language. Python is a computational language designed with the intent to be easily readable and understandable, explicit, simple, and free to use.[1−4] This report is also designed to be self-contained, as all of the necessary data sets, coding, and discussion of analysis methods are contained herein and in the Supporting Information.

The experimental methodology for the laboratory procedure herein is based on the physical chemistry experiment titled "Binary Liquid-Phase Diagram" from *Experiments in Physical Chemistry*,[5] though it has been modified to make use of gas chromatography (GC) in order to determine the mole fraction of the distillate and to use acetone/cyclohexane solutions instead of cyclohexanone/tetrachloroethane mixtures, as tetrachloroethane is reported to be a likely carcinogen.[6] In brief, the experiment requires students in part to analyze aliquots obtained through the fractional distillation of various mixtures of acetone and cyclohexane. The results derived from GC analysis are then used to derive the mole fractions of the liquid and the distillate. With the mole fractions determined,

Received: August 30, 2019
Revised: February 12, 2020
Published: March 4, 2020



https://dx.doi.org/10.1021/acs.jchemed.9b00818
J. Chem. Educ. 2020, 97, 1172−1175

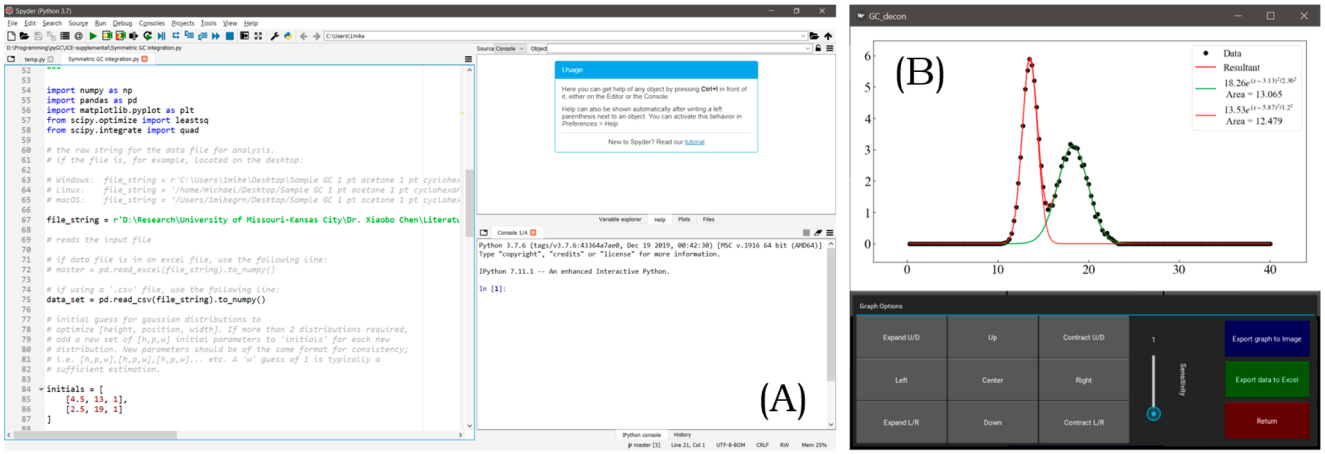


Figure 1. (A) Spyder IDE with the "Symmetric Gaussian.py" executable. (B) The GC_decon GUI main page.

the combination of the derived mole fractions and measured temperatures allows students to construct the liquid−vapor phase diagram.

The core of the experiment is combining their understanding of the theory as derived from Raoult's law[5] with the experimental data determined from GC in a manner that allows them to elucidate the final form of the phase diagram. However, the explanation in the text on how students should determine the mole fraction from the data is lacking in thoroughness and practicality. GC analysis results in a signal versus time data set of strict (x, y) points that require data processing. The particularly difficult nature of the problem arises from the necessity of deconvoluting the data system. Furthermore, an examination of the literature from this *Journal* shows that a suitable method of regression analysis has yet to be demonstrated, particularly with respect to a Python implementation. This report offers such an approach.

METHOD

Fully implementable Python scripts used for data analysis are included in the Supporting Information, as well as example GC data files in .csv format, which is the output file format for the GC unit in the physical chemistry laboratory at the University of Missouri—Kansas City. All of the necessary developing environments and libraries for executing the .py script files are free, available most easily through the Anaconda distribution platform, which can be downloaded freely (downloading instructions for all operating systems can be found in the Supporting Information). Included in the vanilla download is the interactive development environment (IDE) named Spyder, through which the provided Python code can be edited and executed, as shown in Figure 1A. If easy internet access is available, these tools are available for free use in the Google cloud. Also included is a free-standing graphical user interface (GUI)-wrapped application, shown in Figure 1B, for use if teachers simply prefer to avoid interacting with the source code; installation instructions are included in the Supporting Information. All of the necessary information to get started is included in the Supporting Information. To the best of our knowledge, there is no free and widely available software package that allows users to perform the technical capabilities provided herein by either the stand-alone application or the provided Python scripts. It should be noted, however, that we believe the goal here should be to expose students to the technical process of programming. We believe that it is to the students' benefit to experience first-hand how to use data-driven methods to solve the broadest set of problems, beyond just experimental GC deconvolution. For example, curve fitting is a technique utilized in many instrumental techniques, including X-ray diffraction,[7−9] X-ray photoelectron spectroscopy,[9,10] Fourier transform infrared spectroscopy,[11] Raman spectroscopy,[12] and so on, and could in principle be applied to more complex systems such as polyprotic acid/base titrations[13] with the proper modifications.

PEDAGOGICAL FRAMEWORK

The crux of the data analysis for this experiment is understanding that the normal distribution of phase velocity in gas chromatography is generally Gaussian in nature, as the relative transfer from the stationary phase to the mobile phase for any individual molecule within the GC column is random.[14] Trivial methods for determining the integral values of the data set can be done numerically via methods such as the trapezoid rule.[15] However, the error associated with such numerical integration techniques becomes nontrivial and unavoidable if either the interval size of the instrument is sufficiently large or the material signals represented in the data overlap. If there is overlap of the distribution curves, the data should be deconvoluted. As the signal response from the instrument is a superimposed set of Gaussian distributions for the given material identities, data resultants from gas chromatography can be fit to a set of Gaussian functions, as demonstrated in eq 1:

$$I(x) = \sum_{i=1}^{n} A_i \cdot e^{-(x-B_i)^2/C_i^2} \quad (1)$$

where n is the number of material identities being separated by the GC column and A, B, and C are fitting parameters.

For a typical Gaussian distribution function, the parameters A, B, and C define the amplitude, position, and breadth of the distribution function. Determining these parameters allows for the computation of the area associated with the distribution, which then ultimately allows for the determination of material quantity.[14] A multitude of algorithms exist in the literature that take a general-form function with a set of initial parameters and, through least-squares regression analysis, determine the function constants by minimizing the error associated with deviation in the data set from the Gaussian curves. Herein, the Levenberg−Marquardt algorithm[16] is utilized in the Python

https://dx.doi.org/10.1021/acs.jchemed.9b00818
J. Chem. Educ. 2020, 97, 1172−1175

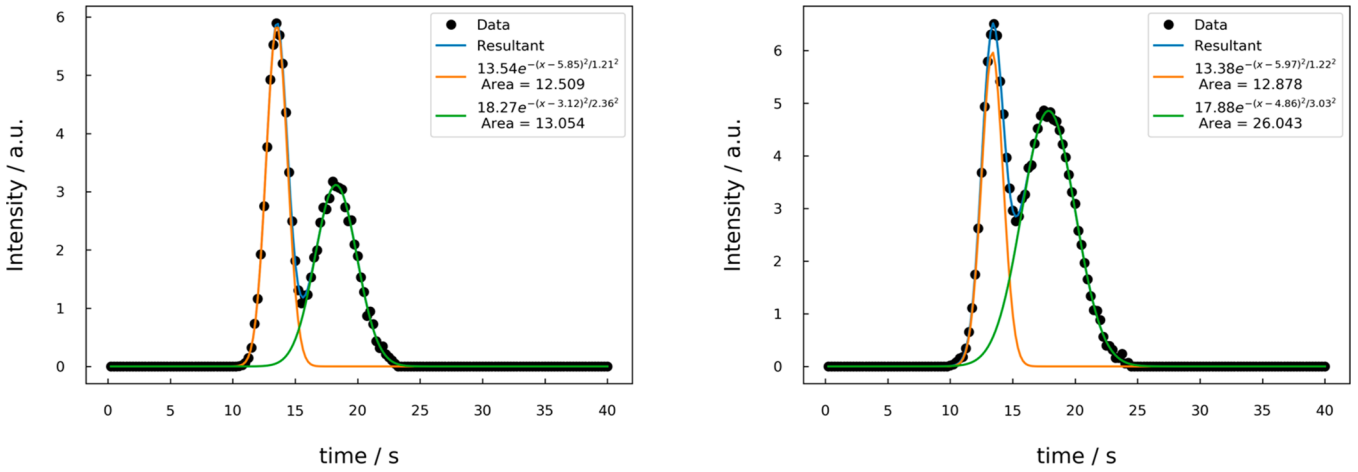


Figure 2. Derived distribution curves from the example GC data files at (a) 1:1 and (b) 2:1 molar ratios of acetone to cyclohexane.

script provided in the Supporting Information (methodological understanding of the algorithm can be extracted from its original publication). Such an approach allows deconvolution of the data sets for a more accurate analysis of the integral areas associated with the signals intrinsic to the material entities. With the determination of the distribution constants, the function can be accurately integrated, thus rendering the relative material quantities determined (Figure 2). For data sets that demonstrate asymmetry due to column overload or other column effects,[17] deconvolution with an asymmetric Gaussian function can be utilized to accurately minimize the residual with respect to the data. In such cases, the Gaussian distribution is incorporated with an error function parameter that allows the distribution to skew toward the direction of asymmetry. Both the included .exe program and .py scripts allow for such asymmetric processing.

IMPLEMENTATION

The program methodology presented herein has been beta-tested by undergraduate physical chemistry students at the University of Missouri—Kansas City using data extracted from the same GC utilized in the teaching laboratory. The Python programs allowed for quick, decisive computation of the calibration curves and the aliquots derived from experiment, resulting in the determination of the molar concentrations of the samples produced via fractional distillation. In general, when the data sets were analyzed with both the trapezoidal rule and the functionalization approach, it was consistently demonstrated that sample analysis had a tighter correlation with the calibration curve derived from the 1:1 and 2:1 molar ratio measurements when this functionalization method was used. This makes intuitive sense, as the error associated with truncating the overlapped portions of the distribution curve is minimized by passing through the data a set of proper distribution functions. The calibration curves for the 1:1 and 2:1 molar ratios of acetone to cyclohexane are included in the Supporting Information for practice (though to be explicit, these curves should not be used as general calibration curves, as the retention time and intensity response are highly dependent on the instrument parameters). For students who were interested in learning how to use the resources presented to solve more problems outsize the strict confines of the liquid−vapor phase diagram experiment, we would typically demonstrate that this data processing methodology of least-squares regression analysis can be applied to data sets and functions beyond Gaussian distributions and GC, such as pseudo-Voigt functions for modeling particle size in X-ray crystallography[18] and exponential functions for modeling the kinetics of photocatalytic systems.[19] The only changes necessary to the Python script are the input data, the initial parameters, and the fitting function. Students seemed to appreciate this utility. Finally, students who used the implementation were given postimplementation surveys for feedback. Though two-thirds of reporting students had no previous experience with Python, a majority of students suggested that they took away from the experience a general understanding of how the problem of GC deconvolution was solved and, more importantly, that they felt they could use the strategies discussed herein to solve similar problems.

CONCLUSION

A more complete methodology for functionalizing gas chromatography data is discussed, with a focus on implementation in undergraduate education in order to introduce students to the idea of utilizing computational devices to functionalize data. The benefit of utilizing this approach is that it provides students access to a set of tools that are free to use, limitless in application, and simple in design and structure. Oftentimes when one is learning a new technology, the most difficult step in beginning to develop fluency in application is the initial step into the domain; utilizing this simple program for a practical purpose such as GC data analysis can be an effective way to provide such a step. From here, the continued use of computers in chemistry can naturally progress.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available at https://pubs.acs.org/doi/10.1021/acs.jchemed.9b00818.

Supporting files: Asymmetric GC integration.py; Symmetric GC integration.py; Sample GC data 2 pt acetone 1 pt cyclohexane.csv; Sample GC data 1 pt acetone 1 pt cyclohexane.csv; Python Installation Instructions.docx; GC_decon Installation and Operation Instructions.docx; Post implementation survey.docx; and JCE jupyter notebook.ipynb (ZIP)

https://dx.doi.org/10.1021/acs.jchemed.9b00818
J. Chem. Educ. 2020, 97, 1172−1175

AUTHOR INFORMATION

Corresponding Author

Xiaobo Chen − *Department of Chemistry, University of Missouri—Kansas City, Kansas City, Missouri 64110, United States*; orcid.org/0000-0002-7127-4813; Email: chenxiaobo@umkc.edu

Author

Michael Green − *Department of Chemistry, University of Missouri—Kansas City, Kansas City, Missouri 64110, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jchemed.9b00818

Notes

The authors declare no competing financial interest.

This Technology Report also has an associated GitHub repository, where the GUI application "pyGC" can be downloaded and Jupyter notebooks tethered to Binder runtimes and the Google Colab cloud computing environment can be accessed. Both Binder and Colab allow initialization of the notebook directly through the Internet browser without any necessary installations. The GitHub repository can be found at https://github.com/1mikegrn/pyGC.

ACKNOWLEDGMENTS

M.G. and X.C. appreciate the support from the U.S. National Science Foundation (DMR-1609061) and the School of Biological and Chemical Sciences, University of Missouri—Kansas City.

REFERENCES

(1) Srnec, M. N.; Upadhyay, S.; Madura, J. D. A Python Program for Solving Schrödinger's Equation in Undergraduate Physical Chemistry. *J. Chem. Educ.* **2017**, *94* (6), 813−815.

(2) Weiss, C. J. Scientific Computing for Chemists: An Undergraduate Course in Simulations, Data Processing, and Visualization. *J. Chem. Educ.* **2017**, *94* (5), 592−597.

(3) Srnec, M. N.; Upadhyay, S.; Madura, J. D. Teaching Reciprocal Space to Undergraduates via Theory and Code Components of an IPython Notebook. *J. Chem. Educ.* **2016**, *93* (12), 2106−2109.

(4) Shi, C. B.; Borchardt, T. JRgui: A Python Program of Joback and Reid Method. *ACS Omega* **2017**, *2* (12), 8682−8688.

(5) Garland, C.; Nibler, J.; Shoemaker, D. Experiment 14: Binary Liquid−Vapor Phase Diagram. In *Experiments in Physical Chemistry*; McGraw-Hill Education: New York, 2009; pp 207−215.

(6) Chen, W.; Zhu, D.; Zheng, S.; Chen, W. Catalytic Effects of Functionalized Carbon Nanotubes on Dehydrochlorination of 1,1,2,2-Tetrachloroethane. *Environ. Sci. Technol.* **2014**, *48* (7), 3856−3863.

(7) Stojilovic, N. E.; Isaacs, D. Inquiry-Based Experiment with Powder XRD and FeS2 Crystal: "Discovering" the (400) Peak. *J. Chem. Educ.* **2019**, *96* (7), 1449−1452.

(8) Green, M.; Tian, L.; Xiang, P.; Murowchick, J.; Tan, X.; Chen, X. Co_2P Nanoparticles for Microwave Absorption. *Mater. Today Nano* **2018**, *1*, 1−7.

(9) Green, M.; Xiang, P.; Liu, Z.; Murowchick, J.; Tan, X.; Huang, F.; Chen, X. Microwave Absorption of Aluminum/Hydrogen Treated Titanium Dioxide Nanoparticles. *J. Mater.* **2019**, *5* (1), 133−146.

(10) Lawrie, G.; Keen, I.; Drew, B.; Chandler-Temple, A.; Rintoul, L.; Fredericks, P.; Grøndahl, L. Interactions between Alginate and Chitosan Biopolymers Characterized Using FTIR and XPS. *Biomacromolecules* **2007**, *8* (8), 2533−2541.

(11) Peterson, K. I.; Pullman, D. P. Determining the Structure of Oxalate Anion Using Infrared and Raman Spectroscopy Coupled with Gaussian Calculations. *J. Chem. Educ.* **2016**, *93* (6), 1130−1133.

(12) Weaver, G. C.; Norrod, K. Surface Enhanced Raman Spectroscopy: A Novel Physical Chemistry Experiment for the Undergraduate Laboratory. *J. Chem. Educ.* **1998**, *75* (5), 621.

(13) Morales, D. A. Mathematical Modeling of Titration Curves. *J. Chemom.* **2002**, *16* (5), 247−260.

(14) McNair, H. M.; Miller, J. M. *Basic Gas Chromatography*, 2nd ed.; John Wiley & Sons: Hoboken, NJ, 2008.

(15) Stewart, J. *Calculus: Early Transcendentals*, 8th ed.; Cengage Learning: Boston, 2014.

(16) Levenberg, K. A Method for the Solution of Certain Non-Linear Problems in Least Squares. *Q. Appl. Math.* **1944**, *2*, 164−168.

(17) Pápai, Z.; Pap, T. L. Analysis of Peak Asymmetry in Chromatography. *J. Chromatogr. A* **2002**, *953*, 1−2.

(18) Green, M.; Liu, Z.; Smedley, R.; Nawaz, H.; Li, X.; Huang, F.; Chen, X. Graphitic Carbon Nitride Nanosheets for Microwave Absorption. *Mater. Today Phys.* **2018**, *5*, 78−86.

(19) Liu, Y.; Li, Z.; Green, M.; Just, M.; Li, Y. Y.; Chen, X. Titanium Dioxide Nanomaterials for Photocatalysis. *J. Phys. D. Appl. Phys.* **2017**, *50* (19), 193003.