Submitted to: J. Chem. Ed.

An Instrument Assembly and Data Science Lab for Early Undergraduate Education

Alison Wallum,[†,*] Zetai Liu,[†] Joy Lee,[†] Subhojyoti Chatterjee,[‡] Lawrence Tauzin,[‡] Christopher D. Barr,[+] Amberle Browne,[§] Christy F. Landes,[‡,&] Amy L. Nicely,[§,*] and Martin Gruebele[†,^,*]

[†]*Department of Chemistry and Beckman Institute for Advanced Science and Technology, and University of Illinois at Urbana-Champaign, Illinois 61801, United States*

[‡]*Department of Chemistry, Rice University, Houston, Texas 77005, United States*

[+]*Office of Research, Rice University, Houston, Texas 77005, United States*

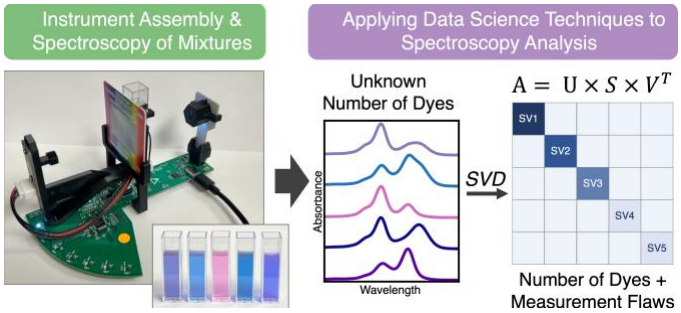
[§]*Danville Area Community College, 2000 East Main Street Danville, IL 61832, United States*

[&]*Department of Electrical and Computer Engineering, and Department of Chemical and Biomolecular Engineering, Rice University, Houston, Texas 77005, United States*

[^]*Center for Biophysics and Quantitative Biology, C-I College of Medicine, and Department of Physics, University of Illinois at Urbana Champaign, Illinois 61801, United States*

Corresponding author emails: Alison Wallum, *awallum2@illinois.edu*; Amy Nicely, *anicely@dacc.edu*; Martin Gruebele, *mgruebel@illinois.edu*

ABSTRACT: As data science and instrumentation become key practices in common careers ranging from medicine to agriscience, chemistry as a core introductory course must introduce such topics to students early and at an accessible level. Advanced data acquisition and data science generally requires expensive precision instrumentation and massive computation, often out-of-reach for even upper-level undergraduate laboratory courses. At the same time, a new generation of affordable do-it-yourself instruments presents an opportunity for incorporation of curricula focused on instrument design and computation into freshman-level courses. We present a new lab for integration into existing courses that starts with hands-on spectrometer building, moves to data collection, and finally introduces an advanced data science technique, singular value decomposition, at an appropriate level with minimal computing requirements. The hardware and software used are modular and inexpensive. The lab was tested in three community college general chemistry sections over two semesters. Previously, students taking these courses did not typically see advanced quantitative chemistry curricula before deciding whether to pursue a bachelor's degree. This lab allowed students to practice data collection and organization skills, use pre-written Jupyter notebooks that perform advanced data analysis, and gain presentation skills. A multi-wave assessment completed by students highlights both successes and difficulties associated with incorporating multiple advanced topics involving data collection and analysis techniques in a single lab.



Keywords: singular value decomposition, spectrometer; community college; Python; food dye; UV-vis

■ Introduction

The modern workforce and academic research have become more intensively data science-driven, requiring skills ranging from fluency in spreadsheets at the most basic level[1] to Python programming or high-throughput data science tools.[2,3] Upper division chemistry curricula have responded with coursework in instrumentation and advanced data analysis,[4,5] often associated with an analytical,[6] physical,[7] or biochemistry lab.[8] With advanced analysis and computational skills no longer considered optional, there is a growing need to find affordable ways to introduce these skills earlier than the junior and senior year in college.[9,10]

By the time most students start to see a more in-depth connection between data science and instrumentation in areas like spectroscopy, they may have already fulfilled their chemistry requirements. One reason such topics are not introduced earlier is that instrumentation can be expensive to purchase, maintain, and update, and advanced analysis techniques can be computationally complex. It is a challenge for upper division instructors to dedicate enough time in a single course to both foster students' instrumentation skills and students' data science skills for the first time in their chemistry education.

Figure 1. An introduction to spectroscopy instruments and data science. This lab focuses on introducing three primary elements in general chemistry curricula, including: (1) allowing students to put together their own instrument from components, (2) practicing data collection on complex unknowns, and (3) implementing data science techniques to interpret scientific data.

The challenge of combining instrumentation and data science is notable in community college settings, where many students prepare to transfer to four-year institutions.[11] Upper-division coursework is not offered at community colleges, and funding levels can make expensive instruments prohibitive. Making the connection between instruments, measurements, and sophisticated data science is thus a difficult task for instructors and students in this context. The community college environment uniquely motivates ways to introduce these topics prior to upper-division coursework, providing students with an early experience surrounding scientific and computational skills that offers flexibility in majors and, regardless of whether or not they transfer

to a four-year institution, career choices.[12] The idea is also relevant to four-year institutions who are considering 'front-loading' analysis skills often reserved for upper division courses.[13]

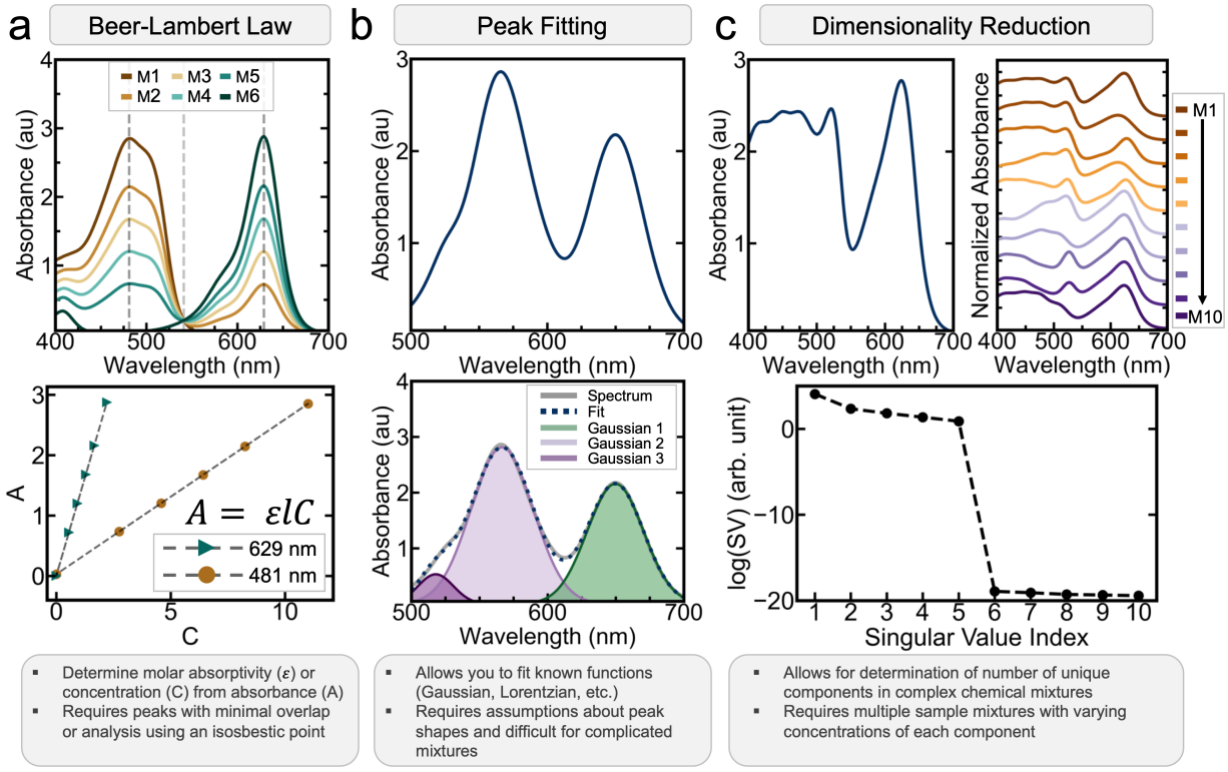


Figure 2. Comparison of three different analyses of absorbance spectra. (a) The Beer-Lambert law allows students to measure absorbance A at a particular wavelength, and if concentration C is known, infer the molar absorptivity ε at that wavelength. This is best implemented when peaks are well-separated (outer dashed lines). Only for mixtures with little overlap of absorbance peaks or by observing an isosbestic point (center dashed line) for two components can one "eyeball" the number of dyes in mixtures such as M1 – M6, otherwise counting components becomes impossible. (b) Peak fitting allows students to estimate the number of components in a mixture and chemical parameters such as peak widths by fitting multiple peaks to a function such as a Gaussian. This is problematic for counting components because an individual molecule's absorbance may consist of several peaks, causing one to overestimate the number of components, or overlapping "shoulders" on peaks may be missed, causing undercounting. More complicated peak fitting algorithms can overcome these issues using techniques such as dimensionality reduction. (c) Dimensionality reduction techniques, including SVD, can be applied to analyze absorbance data of complex spectra with many mixed chemicals to determine the number of unique components, given sufficiently many input spectra of mixtures. This technique does not directly give us information about physical chemical parameters such as molar absorptivity, ε. However, it is a powerful technique for determining unique components when very little is known about the mixtures, and can successfully be applied to complicated mixtures like the example shown in (c) with 5 components in 10 random mixtures (indicated by 5 singular values "SV" with a large magnitude).

With this in mind, we developed a lab module that gives students hands-on experience putting together an inexpensive Raspberry Pi-based spectrometer,[14,15] collecting data on complex unknown dye mixtures, and counting the components in a mixture without knowing their individual spectra by using a computational data analysis tool (**Figure 1**). For the data analysis portion, we introduce singular value decomposition (SVD) via spreadsheets and Jupyter notebooks.[3] The module was implemented three times in a community college setting, where access to expensive instruments

and proprietary software can be prohibitive. The module is short enough to be integrated into existing curricula. We conclude with a multi-wave assessment that highlights some of the successes as well as challenges of an early-curriculum instrument and data analysis-focused lab.

■ Lab motivation and timeline

This lab allows students to see first-hand what the inside of a spectrometer looks like, and how programming and data science techniques can extract hidden information from chemical measurements. All three aspects of the lab are hands-on: two experimentally, and the last with active analysis exercises that go beyond typical freshman chemistry experiences. The three activities walk students through the process of building and applying a spectrometer and data science techniques to answer the question: How many unique dye molecules were used to prepare a series of colorful unknown mixtures?

The core structure of the lab, discussed in detail in the next section, is as follows: students assemble basic components of an absorption spectrometer kit (Trimontana, Inc.) into a functioning instrument, and collect data on known food dyes to test their instrument. They then collect absorbance spectra of unknown mixtures of more than two dyes, yielding a series of complex multi-peaked spectra. Finally, they transfer data using a spreadsheet (Excel, Microsoft, Inc.) and use a pre-written Jupyter notebook to determine the number of dyes in the unknown mixtures using SVD. These results are then compared to (1) data collected on a commercial instrument and (2) "perfect" data produced artificially, allowing students and instructors to discuss how advanced data science techniques can be applied to analyze data from imperfect systems like their hand-built spectrometers.

SVD was chosen to show students how modern data science techniques are applied to answer questions in chemistry, as it is a fundamental data science technique and broadly applicable in the sciences.[16] Our basic question for this lab is: How many components are in a mixture? Different approaches to analyze spectra of chemical mixtures are outlined for students as shown in **Figure 2**. A classic approach taught in general chemistry is the Beer-Lambert law in **Figure 2a**. For simple mixtures with enough separation between all component absorbance peaks, analysis of mixtures is often applied to determine either the extinction coefficients of dyes (ε) at a particular wavelength or concentrations (C) of dyes in solution. In this case, information about the spectra and either ε or C must be known beforehand. When this information about the individual components is not

known or absorbance peaks overlap, peak fitting in **Figure 2b** is another option. Usually, a sum of known functions (e.g., Gaussian or Lorentzian) is fit to spectra, but this can easily miss components (a small shoulder on a big peak) or add spurious components (a dye that has two peaks would be counted as two separate molecules, like the one shown between 500 – 625 nm in the bottom of Figure 2b). These challenges can be overcome by applying linear dimensionality reduction techniques such as SVD, which assume only that the data is a sum of independent components. As discussed in detail in section IV, SVD identifies unique components by decomposing the absorbance spectral data into three matrices: the data in **Figure 2c** form the input matrix for SVD, with wavelength along the horizontal axis and mixture number along the vertical axis; one of the matrices output by SVD contains the singular values 'SV' that rank components by significance (5 singular values are significant in **Figure 2c**, thus there are 5 components in the mixture).

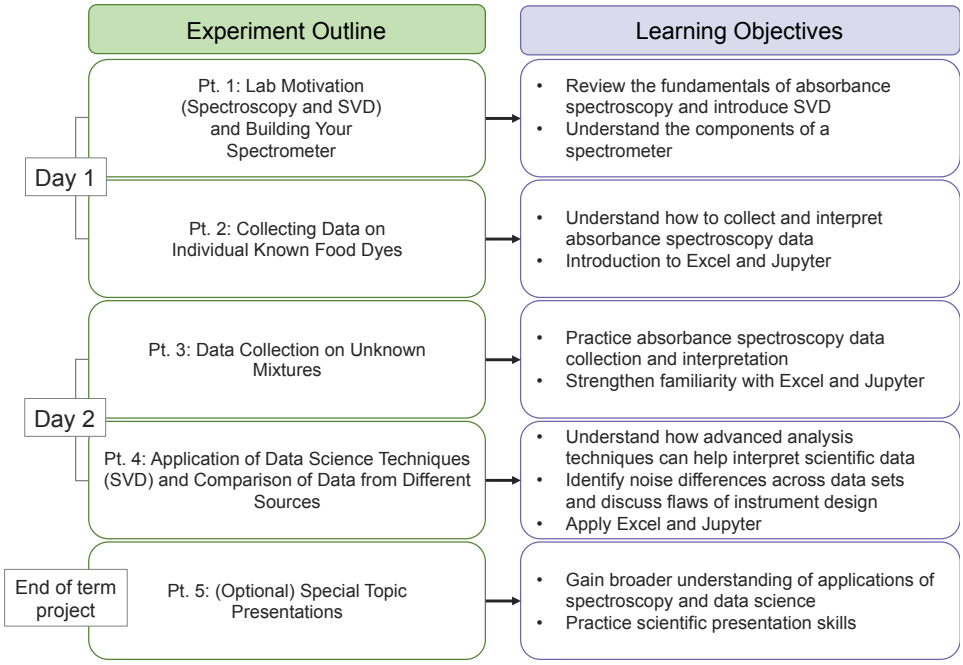


Figure 3. Suggested experiment timeline and learning objectives for "How many unique dye molecules were used to prepare a series of colorful unknown mixtures?". In the spring and fall, the experiment portion of the lab was implemented during additional lab time over the course of three days or over the course of two lab classes respectively. This translates to four traditional laboratory hours with time post-lab to complete the analysis if needed. In addition to this, both the Fall and Spring implementations included a class at the end of the semester focused on group presentations.

We aim to show students how a simple answer (how many different dyes are in a set of unknown mixtures), hidden when one just looks at complicated raw spectra, emerges from advanced data analysis techniques like SVD. They additionally see how a few dyes produce many colors when combined, drawing a connection with how a food scientist might apply reverse engineering to

deduce combinations of food dyes used to make different Jell-O products, or how a coatings chemist may determine the number of pigments used in a given set of paints.

For our implementation at Danville Area Community College, first-year undergraduate students in three general chemistry sections completed this lab either in three short sessions over a three-week period (Spring 2022) or in two lab periods (Fall 2022). Our module followed another spectroscopy-focused module earlier in the semester that introduced the Beer-Lambert law using a commercial 'black box' UV-Vis spectrometer. Students worked in groups of two to three for all portions of the lab and submitted final reports and projects together. Roughly ~57% of students were pursuing an Associates in Science (focuses reported included biology, nursing, engineering, general science, veterinary science, horticulture and agriculture), ~25% were pursuing an Associates in Art, ~6% were pursuing an Associates in Business, and the remainder were undeclared. Most incoming students had taken chemistry in high school; however, the majority of students did not have an introduction to physics and calculus in high school.

This lab includes several computational and analysis-focused topics that are likely new to students. For this reason, we structured the lab for students who have had only a basic introduction to spectroscopy, either in an earlier semester or in their current lab (as was the case here). The first two sessions of the lab allowed students to carry out Parts 1 – 4 in **Figure 3**. An additional presentation portion of our lab was implemented as an end-of-semester project, where students chose a topic relevant to a portion of the lab (e.g., applications of spectroscopy, data science, absorbance and food dyes) and gave 5-to-10 minute presentations. This allowed students to hone presentation skills at the conclusion of the lab. If instructors have time before the end of the semester, the basic lab can be extended with the group presentation during a third lab period.

This lab can be structured as a 4-6 hour activity, tailored to fit into a university's current course structure (**Figure 3**). Depending on how familiar students are with absorbance spectroscopy and spreadsheet manipulation, the spectrometer building and data collection can be accomplished during either one or two labs (i.e., 1-2 days in a lab course), with some time left for students to begin the analysis and plotting their data in the spreadsheet. Total cost and lab-specific supplies are included in the instructor guide document on GitHub, in addition to instructions for preparing materials for the lab.

We suggest that instructors provide a pre-lab lecture introducing singular value decomposition and reviewing basic absorbance spectroscopy, covering topics similar to those outlined in Figure

2a. Assuming students have little experience with spreadsheets and Jupyter, we also recommend a pre-lab assignment and/or a short guided activity with the instructor during a lab lecture, where challenging concepts are outlined (e.g., using a spreadsheet, navigating a Jupyter notebook). Examples of lecture materials and pre-lab assignments to support the instructor can be found on the CAFF GitHub.[17]

Given the challenge of assembling a spectrometer, collecting data, and doing sophisticated analysis in a short period of time, our implementation of the lab focused on students applying data science using spread sheets and Jupyter notebooks, rather than coding scripts themselves. Following along with pre-written code and seeing what syntax is used to give a certain output is an effective way for many people to start learning about programming,[18] and we hope that this will serve as a preface for later courses that focus on a deeper understanding of coding and developing programming skills. Ideally, this early exposure to programming and data science helps reduce the initial learning anxiety surrounding these topics in advanced courses, providing a smoother transition to upper-division coursework.

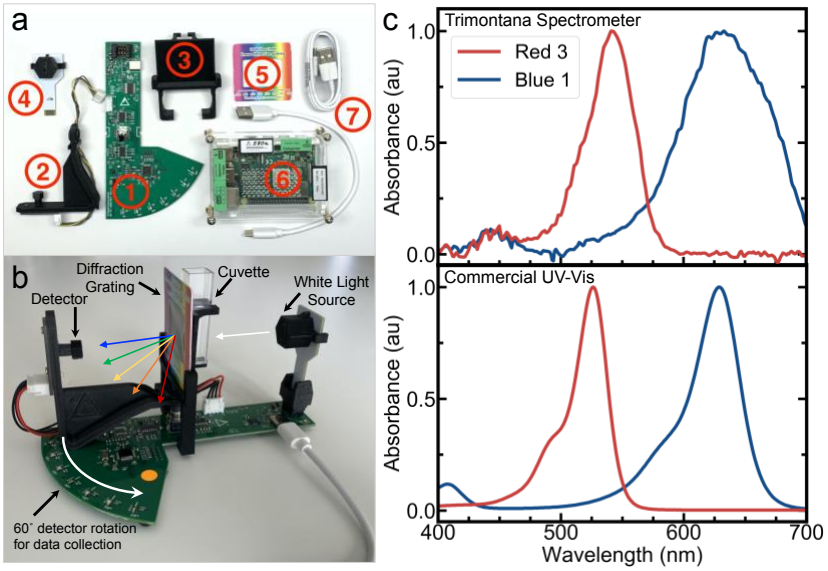


Figure 4. Trimontana spectrometer kits for absorbance spectroscopy experiments. (a) Each Trimontana spectrometer kit includes (1) a PCB board, (2) swivel arm detector, (3) grating and cuvette mount, (4) white light source, (5) diffraction grating, (7) cables, and (8) a Raspberry or Orange Pi. (b) Constructed spectrometers operate by passing the white light source through a sample and a diffraction grating. The swivel arm detector is then manually rotated by 60˚, and the intensity of light is collected at each angle to yield an absorbance spectrum. (c) Absorbance spectra collected on a Trimontana spectrometer of food dyes Red 3 and Blue 1. Data collected on a Trimontana spectrometer (top) show similar spectra when compared to commercial UV-Vis spectrometer data (bottom), with peaks slightly shifted and broadened due to a wavelength accuracy of ±10%, as well as reduced blue light throughput.

■ Implementation

Part I: Lab Motivation (SVD and Spectroscopy) and Building a Spectrometer. On day one of the lab (first session, Parts I and II), we recommend instructors begin with a pre-lab lecture to provide motivation and context for the different portions of the lab. This should include a brief review of absorbance spectroscopy followed by an introduction to how singular value decomposition works and will be applied.[17] Students then build their own spectrometers and begin to discuss the relevance of the different components needed to collect spectroscopy data. Previous work has emphasized the benefits of student-built spectrometers in chemistry education including improved student confidence in instrument operation and improvements on students understanding of instrument troubleshooting.[4] Other studies also reported improved understanding of key concepts including the spectrometer components, the difference between transmittance and absorbance, and the Beer-Lambert law.[19] Additionally, meta-analyses of hundreds of individual studies of STEM courses strongly supports that active learning strategies (such as building a spectrometer as part of the exercise) improve student outcomes.[20] Finally, given the data science focus of this lab, using a hand-built instrument instead of a commercial spectrometer allows for discussions surrounding noise reduction using SVD and analysis of imperfect data, primarily during the pre-lab lecture and analysis portion of this lab.

For the building activity, we purchased spectrometer kits from the educational instrument company Trimontana to provide students with all the necessary components required to assemble their own manually operated absorption spectrometers. **Figure 4a** shows the components of a spectrometer kit and **Figure 4b** shows a constructed spectrometer. Spectrometers operate by passing a light from an LED white light source through a liquid sample in a plastic cuvette, then out of a slit and through a transmission diffraction grating. A swivel arm detector is manually rotated by the student from 0˚ to 60˚, and the intensity of the transmitted diffracted light is collected by a detector, each angle corresponding to a specific wavelength. The relationship between transmitted light and absorbed light can then be used to output an absorbance measurement. Spectrometers are set up so that they can operate offline through a Raspberry Pi or Orange Pi, or online by connecting the spectrometer directly to a computer and connecting to the course's JupyterHub (**SI Figure 1**).

Part II: Collecting data on individual known dyes and building familiarity with data analysis. Once the spectrometers are built, students collect absorbance data of water (blank) and solutions of at least two known food dyes. Skills learned while students test their spectrometers include making sure not to block the detector or bump the diffraction grating mid-measurement. This step also gives students the opportunity to compare their homebuilt spectrometer data with commercial UV-Vis data (Shimadzu UV-1800) (**Figure 4c**), which is provided or readily available online for the dyes used in this experiment. After building and testing their spectrometers, groups are prompted with questions about the spectrometer components and how absorbance is calculated from transmitted light. They are then asked to list and discuss two or three steps in data collection that they think are important for collecting high quality data, and how someone might either introduce error or improve their measurements by modifying these steps. Before students begin Parts III and IV, we hope to highlight how measurement imperfections relate to features of the data they collected, such as peak positions and line widths. This enables a broader discussion of data quality in chemical measurements.

This section of the lab also allows for a brief introduction to using spreadsheets and Jupyter notebooks. Spreadsheet skills translate to a variety of applications both in and outside of the sciences[21] and the software is often available through an educational license. As a recommended option, the instructor can demonstrate relevant spreadsheet operations as a way to organize and save data during a pre-lab lecture.[17] Students will ideally have access to computers (available as part of instructional support at many colleges) and can follow along during the pre-lab lecture on spreadsheet use, making this a hands-on demo-lecture to reduce barriers and anxiety about using numerical software for the first time during a lab session.

Jupyter notebooks are introduced as the interface for operating Trimontana spectrometers. The graphical user interface (**SI Figure 1**), allows students to first see what the Jupyter environment looks like and start using it before interacting directly with any code in Part IV. Data collected through the Trimontana Jupyter environment can be easily downloaded as a CSV (comma-separated values) file for additional manipulation. This step connects data collection and what students learned about spreadsheet manipulation by having them organize and plot their water and known food dye data.

Part III: Absorbance data collection for unknown mixtures. On day 2 of this lab (parts III and IV), students use their spectrometers followed by SVD to determine the unknown number of unique dyes in a collection of colorful mixtures. They are not provided with spectra of the individual unknowns, only five mixtures. This portion of the lab is framed in the context of a food scientist attempting to reverse-engineer colorful food samples and finally answer the question: how many dyes were used to make the five samples? Students are provided with a thorough outline of the workflow for this portion of the lab in their laboratory manual,[17] and a brief outline from the course instructor before beginning.

First, students are given a collection of five unknown mixtures and asked to collect two absorbance spectra for each of them. For our implementation, all mixtures contained combinations of Red Dye 3 (Flinn Scientific), Green Dye 3 (Flinn Scientific), and sulforhodamine 101 (Sigma-Aldrich). These dyes were chosen because they are safe and absorb in the 500 – 700 nm range where the sensitivity of the Trimontana spectrometers is highest (**SI Figure 2**). To ensure consistent data collection, after students collect two spectra for each mixture, they are prompted to collect a third replicate if the two initial measurements for a single mixture look significantly different. After student spectrometers are set-up, this portion of the lab takes ~1.5 hours to complete. Example spectra and instructions for preparing mixtures used when developing and testing this lab are included in the instructor guide spreadsheet (Excel document).[17]

Part IV: Applying data science (SVD) to count unknowns in a mixture and discuss data quality. Next, students apply the widely used data science technique, SVD. For both implementations, most students start the analysis in class and submit their results as a post-lab assignment. As a part of the day two pre-lab lecture, we recommend that instructors remind students of the broader workflow and concepts surrounding SVD so that they are prepared to complete the analysis at home if needed.

This technique provides insight into how many separate components make up a data set,[22–24] and can be applied to complicated data sets, which ultimately can be described by a set of fundamental components. For this lab, SVD allows students to identify the number of significant singular values associated with their data, which provides a count of the number of unique dyes in their unknown mixtures.

Figure 5 provides a simplified example of how SVD is implemented with absorbance spectroscopy data. In this example, the data set 'A' consists of $M = 5$ spectra of mixtures, each with data for $N = 6$ wavelengths (**Figure 5a**). The data is turned into a data matrix A with N rows and M columns. The SVD algorithm decomposes the matrix A into a product of three matrices:

(1) The N by M matrix U, which contains N functions that, when added together in the right combination, exactly reconstruct the data in A.

(2) The matrix V^T, which contains the coefficients for those combinations.

(3) Of particular interest here, the matrix S, a diagonal matrix containing $M = 5$ (in our example) singular values 'SV' that express how significant each function in U is.

The bigger an 'SV' value is, the more important its corresponding function. For example, let's say our five spectra were made by mixing just two dyes. The plot of SV in **Figure 5b** will only show two large singular values, indicating two components in this data set (in contrast to the five larger values shown in **Figure 2c**, reflective of the five components). The remaining three singular values for this system should be near or equal to zero as there is little noise in the data, while comparatively noisy data will result in small non-zero singular values.

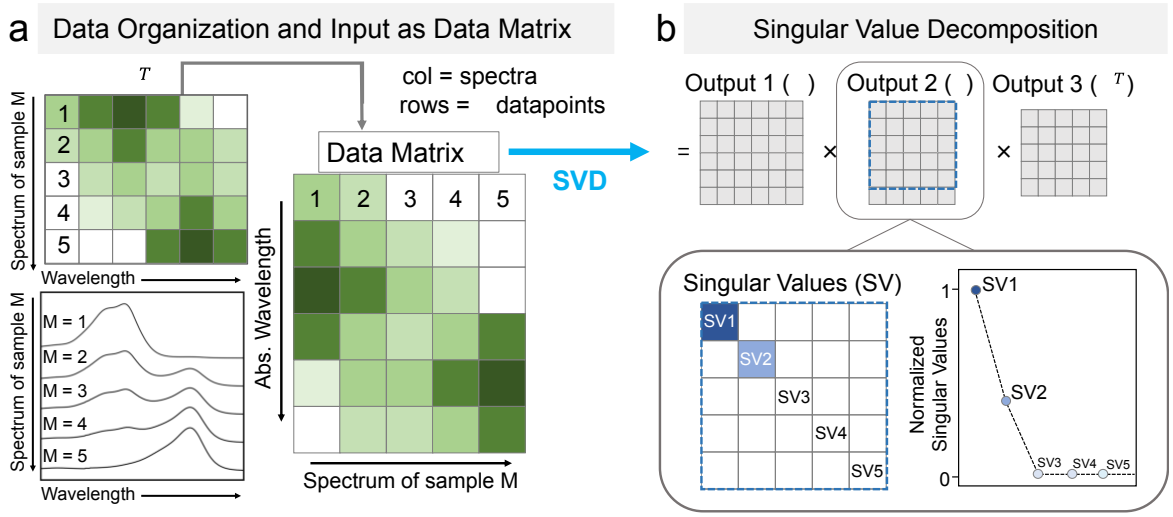


Figure 5. Workflow for SVD analysis. (a) Students organize/format their data for unknown mixture spectra in a spreadsheet for input into a Jupyter notebook script available on their course JupyterHub. To allow for easier data manipulation in Excel, students format their data as the matrix A, with M columns for each spectrum and N rows for each wavelength data point. If this data is organized as A^T (the transpose), it can be visualized similar to a typical wavelength vs. spectrum graph, with dark green points corresponding to data points with large absorbance values (peaks in the spectrum). (b) SVD of data matrix A is performed using the Python library SciPy, giving an output of three matrices. In our explanation of SVD in the SI, these matrices are U, S, and V^T. The diagonal second matrix S contains the singular values, which are printed out and plotted using the available Python script. The example shown here would correspond to a sample with only two large singular values, indicating there are two primary components in the mixtures from (a).

Given this, the number of significant entries in matrix *S* allows us to count unknown components in the spectrum without having to resort to visual-inspection, fitting peaks that may or may not resemble the actual unknown spectra, or other unreliable methods. If there is not a clear break in the singular values, then either there are more components than mixtures we studied, or the noise in the data is high, contributing spurious components. Either way, this alerts us to data quality issues that then deserve further investigation. Thus, SVD can be a useful metric for data content and data quality. The analysis can show students how SVD can be applied to (1) determine the number of different and significant components in their spectra (here, the number of unique absorbance features, each one corresponding to a single dye), and (2) assess noise/error in measurements through comparison of their Trimontana results with higher signal-to-noise ratio data provided by the instructor.

An additional description of the linear algebra behind SVD is included in the Supplementary Information, but the explanation in **Figure 5** should suffice to guide students who have never studied linear algebra. An intuitive verbal explanation for what SVD can tell us is included in the course lecture materials and laboratory manual as well.[17] Similar to other first or second-year chemistry labs that have students applying advanced computational tools to teach chemistry,[25–27] our focus here is to provide students with a functional understanding of what this technique does and how it can be applied to study chemical systems, reserving a complete understanding of mathematical proofs involved for future coursework.

The SVD analysis is performed using a Jupyter notebook, chosen as a programming environment for three reasons. First, running the analysis and organizing course data on a JupyterHub provided continuity with the user interface for the Trimontana spectrometers. Second, the computing environment is free, open source, and widely used to facilitate scientific data sharing and exploration. Third, many recent efforts to incorporate programming in upper division chemistry curricula have also focused on using Python and Jupyter notebooks, allowing continuity with our lab and potential future coursework.[28,29]

Students begin the analysis portion by using a spreadsheet to organize all 10 absorbance spectra into the data matrix *A* and uploading it to their group's JupyterHub directory (**SI Figure 3**). For the SVD analysis a script is available for students, broken down into six Jupyter cells. This allows students to see what Python syntax looks like and execute each cell stepwise, seeing what the output is for each portion of the code (**SI Figure 4**). The first portion of the code imports the necessary

libraries for the script, including `Matplotlib`, `SciPy`, `NumPy`, and `pandas`.[30–33] The second portion allows students to import their data and prints the data so that they can see what was imported. The third portion computes the singular values (using the library function *linalg* from SciPy) and prints out the diagonal elements from the singular value matrix. The next two portions of the script plot the singular values with and without a log scale to help students visually see how many singular values are significantly larger than the rest. Finally, the last cell saves all singular values as a .csv file, allowing students to easily download and save their singular values in a spreadsheet. By allowing students to walk through each step of the code, observing the input and output of each section, we provide students with a basic introduction to the capabilities and structure of analysis code without requiring prior programming knowledge.

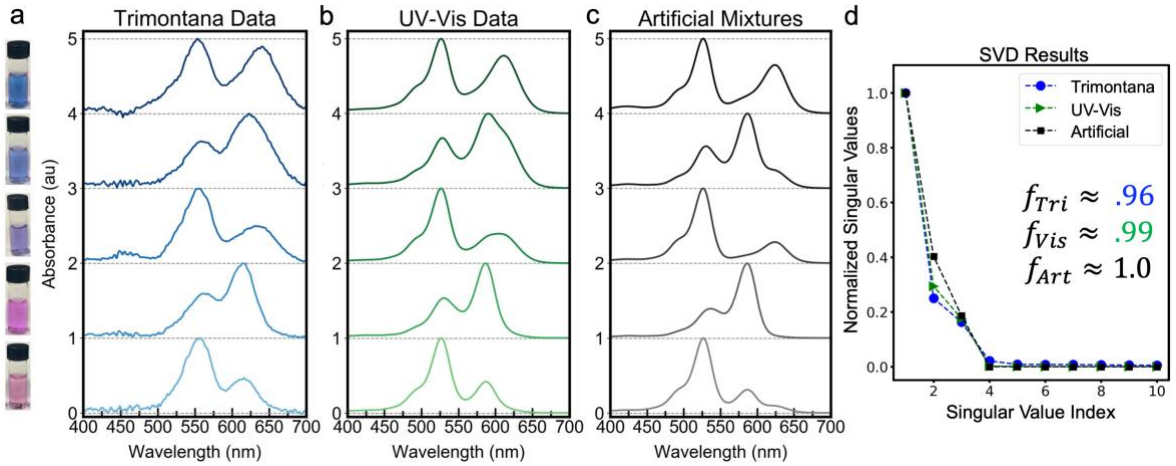


Figure 6. Compiled results for parts III and IV of the lab. (a) Photos of "unknown" mixtures containing combinations of three different dyes and their corresponding absorbance spectra collected on the Trimontana spectrometer. (b) UV-Vis data for the same five mixtures show similar absorbance spectra, with some peaks blue shifted when compared to the Trimontana data. (c) Artificial absorbance spectra of five mixtures made from the same dyes generated from linear combinations of individual dye spectra. (d) Performing SVD on data sets (a), (b) and (c), and plotting the normalized singular value magnitudes shows that in each case, there are three singular values that are larger in magnitude. When we calculate the coefficient f for all three datasets (**Eq. 1**), we find three singular values are needed to describe over 90% of the data. The *x*-axis is the number of singular values, *k,* from *k*=0 to *M*. Values drop in magnitude rapidly after the first three singular values, which describe about ~96% of the data for the Trimontana data (f_{Tri}), and about ~99% for the UV-Vis (f_{Vis}) data. Synthetic data generated by adding combinations of individual dye spectra contains no noise, and therefore the first three singular values account for ~100% of the data (f_{Art}).

After computing the singular values, students use these values to determine the number of dyes in their unknown mixtures. To do this, students compute what fraction *f* of the data is accounted for by $k \leq M$ of the singular values (M = 10 in the example in **Figure 6**, with two replica spectra for each 5 mixtures),

$$f = \frac{\sum_{i=1}^{k} SV_i}{\sum_{i=1}^{M} SV_i}. \quad [1]$$

Thus, as k increases from 0 to $M = 10$ in our example in **Figure 6**, f increases from 0 to 1 (or 100%). Because the hand-assembled spectrometer uses inexpensive and thus imprecise components in comparison to a commercial UV-Vis spectrometer, we define the number of significant singular values as the minimum number of singular values where $f > 0.9$. This value corresponds to selecting the number of components that are needed to describe over 90% of the original dataset collected by students. **Figure 6d** (blue) shows results produced from student data in Figure 6a.

As a final step in their analysis, students are now ready to discuss the effect of data quality on the singular values, as shown in **Figure 6**. They are provided with two additional datasets and asked to repeat the process of finding the number of dyes using SVD. These data sets have spectra for (1) mixtures with the same three dyes collected on a commercial UV-Vis spectrometer and (2) a synthetic data set that contains absorbance spectra for mixtures generated by taking exact linear combinations of individual dye spectra without any noise.

With appropriate data collection on the Trimontana spectrometer, all three data sets allow for the determination of the correct number of dyes, however the amount of data described by the first three singular values increases for the UV-Vis and artificial data (**Figure 6d**). This portion takes advantage of the inherent flaws present in data from hand-built instruments compared to commercial instruments to turn these flaws into useful features of the lab. Students can discuss the differences in their results when the quality of their data changes, and the instructor can (optionally) introduce the application of SVD as a noise filtering tool. As an instructional aid for such a discussion, an example of applying SVD to noise reduction of Trimontana spectrometer data is included in the **Supplementary Information** and on the CAFF Git Hub[17] for instructors to use in their pre- or post- lab lecture materials. For the example shown in **SI Figure 5**, by zeroing out the small singular values in the matrix S and reconstructing the data matrix $A=USV^T$, a noise-reduced spectrum is obtained. The reduction results because small singular values correspond to randomly fluctuating functions in the matrix U that describe the noise in the data. This algorithm is beneficial when compared to applying alternatives such as low pass filters ('smoothing') to the data because it does not cause broadening of the underlying data; it simply removes the components from the data that do not consistently show up with the same shape in every spectrum because they correspond to noise.

Part V: Special Topic Presentations

For the optional final portion of the lab, students worked in teams, previously assigned for data collection, and organized a 5 to 10-minute presentation on a topic surrounding applications of spectroscopy or a related field. In both the fall and spring implementation, groups were given the option to either come up with their own topics or choose from a list of prompts related to the lab. Topic prompts ranged from how spectroscopy and data analysis is applied in different fields, to how dyes and pigments are used in industry. Most students chose projects surrounding applications of spectroscopy in different fields, while some chose to focus specifically on chemical mixtures and detecting their different components. After topics were approved, groups put together formal presentations and presented them at the end of the semester. On the day of the presentations, students were required to fill out review forms for each presentation and ask questions in an effort to make presentations more interactive.

■ Assessment results

Student assessments for this lab were carried out through surveys consisting of questions surrounding student backgrounds, interests in general topics relevant to the lab, and impressions of specific portions of the lab. Surveys were done both for the Spring 2022 semester (1 section) and Fall 2022 semester (2 sections) (see **SI** for the Survey Instruments used). To better gauge students' overall impressions of this lab and the group project, additional questions were added to the Fall 2022 survey. This revised assessment allowed for better evaluation of this specific lab's success in relation to the learning objectives, rather than students' overall impressions of the entire course. Given this and improvements made to the lab structure in Fall 2022, these assessment results are the primary focus of our discussion below. Aggregate results focused on students' assessment of how difficult lab tasks were for both spring and fall surveys are discussed in brief and can be found in the Supporting Information.

Though this lab was challenging for students, student responses indicating that they benefited from the lab point towards an overall positive student experience (**Figure 7**). Following the lab and presentation project, students noted significant improvements in their understanding of spectroscopy, data science, and how these fields are used to solve problems in chemistry. In a Likert scale assessment, ~89% of students reported an increase in their understanding of

spectroscopy and its applications, while 79% reported an increase in their understanding of applications of data science/programming. Students also reported significant improvements in their comfort levels with being able to apply the skills they learned in the future. Improvements in students' comfort with using Excel was the most notable (**Figure 7**), despite students also reporting this being one of the most challenging skills (**SI Figure 6**). This aligns with the fact that students in the Fall 2022 section were provided with robust supplemental materials to help with gaining skills needed for Excel tasks. Students reported the next highest improvement in their comfort with spectroscopy data collection and analysis, followed by the use of Jupyter notebooks. These findings highlight the rewards and challenges of introducing advanced tools in early undergraduate chemistry education.

As the majority of students in the course had little to no experience with the computational tools introduced, students expectedly found the data analysis and data science portion of the lab the most challenging (**SI Figure 6**). The majority of students were nevertheless able to perform the analysis correctly. Of the 11 groups who completed their reports in Fall 2022, 9 collected quality data that would have correctly determined the number of dyes, and 8 groups completed the data science and analysis portion correctly. We found that the instructor who taught the class for a second time in the Fall semester had 100% success in getting correct spectra and analysis from each student group who completed a report, showing that faculty can become experienced with the material fairly quickly. Free response questions from students indicate that one of students' biggest challenges with the analysis was troubleshooting problems with the computational tools used. Students specifically noted challenges with Jupyter in the Fall 2022 section. This underscores the importance of the supplemental exercises and instructions for Excel incorporated in the Fall 2022 implementation and future improvements to the supplemental materials for Jupyter. Students in sections who were able to reference these during their work had better success with completing the analysis correctly.

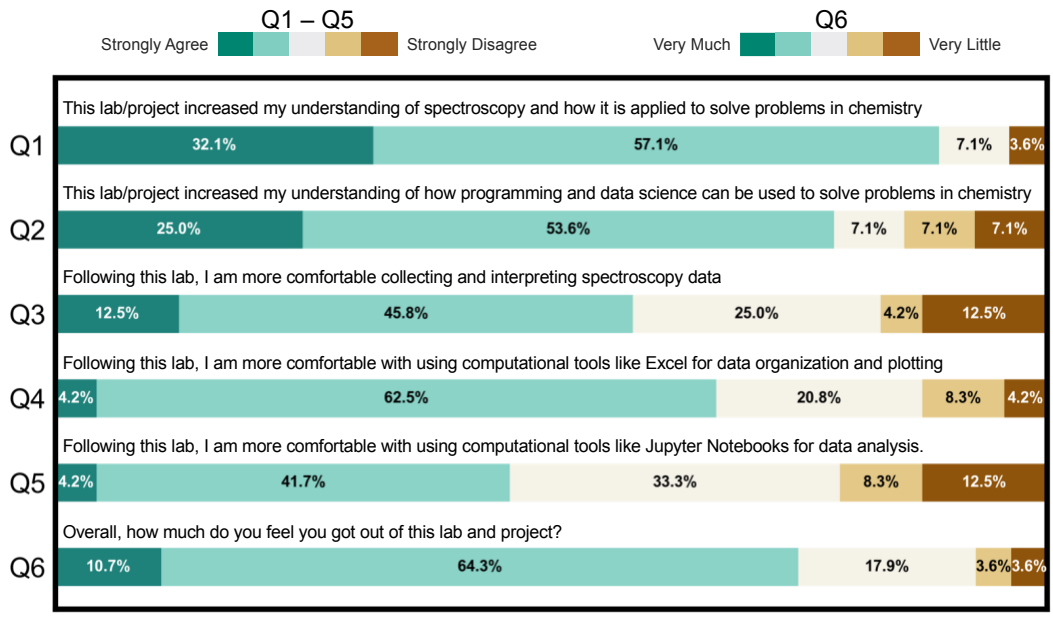


Figure 7. Fall 2022 Likert- scale assessment responses for student impressions of experiment and presentation experiences. Sample size is 28 students for Q1, Q2, Q6 and 24 students for Q3 – Q5.

In the free response section of the survey, many of the students in both the Spring and Fall noted that the building activities and the group presentations were their favorite portions of the lab. In line with previous work surrounding spectrometer building activities,[4,34] many students who favored the building activity noted that they enjoyed constructing an instrument and seeing it work. This alongside the ability to discuss how flaws of the instrument design can be evaluated and overcome using data science techniques highlights the advantages of incorporating a spectrometer building activity into this lab. For the students who reported the presentation portion of the lab as their favorite, many mentioned that they were excited to pick a topic that they found interesting related to lab, and that they enjoyed hearing about the different practical applications of techniques the groups discussed. While this lab can be implemented without the special topic presentations, given the significant number of students who reported the research and presentation portion of the project as their favorite, we highly recommend including this portion in the course structure provided time allows.

For the Spring 2022 implementation, background material for all portions of the lab was implemented in the form of a short overview and heavily relied on written or video instructions for specific information on how to carry out lab tasks. To address student difficulties seen with the quantitative nature of the lab, we developed a set of comprehensive instructor resources for

implementing this lab incorporated in the Fall 2022 implementation.[17] Instructor feedback also stressed that one of the most important aspects of this lab is making sure that there is clear communication and instructor support for (1) how to carry out the analysis and (2) tools for communicating to students what the bigger picture take-aways are for each portion of the lab. We recognize that one of the key challenges of incorporating new curricula surrounding programming and data science is that, given these are modern techniques, we cannot assume all instructors have the resources available to teach these topics or are familiar with them prior to adopting this lab. For future implementations, an instructor guide, detailed instructions, and lecture slides for guided activities and background on the analysis are available on the laboratory GitHub for this course in an effort to enable broad accessibility of this lab.

■ Conclusions and Future Work

Applications of data science and programming skills are becoming ubiquitous in scientific settings[35], and analysis of modern spectroscopy data often requires application of these techniques. Despite the reality that advanced chemical data acquisition is often costly, there are simple optical components and data processing resources that are affordable enough to be accessible to any freshman-level laboratory. In this work, we aimed to emphasize the relationship between instrument design, data acquisition, and computational scripts that extract meaning from data in spectroscopy through a four-part lab with a presentation (**Figure 2**). This lab takes advantage of the documented benefits of using hands-on building activities in labs, with the focus of connecting spectroscopy and data science. Additionally, it uses the inherent flaws of hand-built instruments to facilitate discussions surrounding instrument design, data quality, and how data science techniques can help us analyze imperfect data. Spectrometer building proved to be an enjoyable hands-on start to the lab, easing students into the challenge of applying SVD. Alternative implementations of this lab may expand into a multi-week project throughout the semester rather than a traditional course structure. This could allow students to, for example, write their own analysis scripts, compile data from the entire course, or test out different experimental conditions. Although data science, even at the simplest level of spreadsheets and pre-written analysis tools, are clearly the most challenging item in the lab for students, we believe that introducing these concepts early on will also improve student success when they encounter them again more deeply in upper division courses.

ASSOCIATED CONTENT

■ Supporting Information

Contains a link to additional information at the NSF Center's (CAFF) GitHub, as well as additional figures, explanatory text about SVD, and the content of the survey instrument used.

■ AUTHOR INFORMATION

Corresponding Authors

Alison Wallum − *Department of Chemistry, University of Illinois at Urbana-Champaign, Illinois 61801, United States; Email: awallum2@illinois.edu*

Amy Nicely – *Danville Area Community College, 2000 East Main Street Danville, IL 61832; Email: anicely@dacc.edu*

Martin Gruebele − *Department of Chemistry, University of Illinois at Urbana-Champaign, Illinois 61801, United States; Center for Biophysics and Quantitative Biology, Beckman Institute for Advanced Science and Technology,Carle-Illinois College of Medicine, University of Illinois at Urbana Champaign, Illinois 61801, and Department of Physics, University of Illinois at Urbana Champaign, Illinois 61801, United States; orcid.org/ 0000-0001-9291-8123; Email: mgruebel@illinois.edu*

Authors

Zetai Liu − *Department of Chemistry, University of Illinois at Urbana-Champaign, Illinois 61801, United States; Email: zetail2@illinois.edu*

Joy Lee − *Department of Chemistry, University of Illinois at Urbana-Champaign, Illinois 61801, United States; Email: joylee3@illinois.edu*

Subhojyoti Chatterjee − *Department of Chemistry, Rice University, Houston, Texas 77005, United States; Email: sc168@rice.edu*

Lawrence Tauzin − *Department of Chemistry, Rice University, Houston, Texas 77005, United States; Email: ljt1@rice.edu*

Christopher D. Barr – *Office of Research, Rice University, Houston, Texas 77005, United States; Email: chrisb@rice.edu*

Amberle Browne – *Danville Area Community College, 2000 East Main Street Danville, IL 61832; Email: a.browne@dacc.edu*

Christy F. Landes − *Department of Chemistry, Department of Electrical and Computer Engineering, and Department of Chemical and Biomolecular Engineering, Rice University, Houston, Texas 77005, United States; Email: cflandes@rice.edu*

Acknowledgements

This work was supported by grant CHE 2124983 from the National Science Foundation for the Center for Chemical Innovation, "Center for Adapting Flaws into Features" (CAFF). A.W. is supported by an NSF Graduate Research Fellowship (grant DGE 1746047). The authors would like to thank Prof. Jay Deiner of City University of New York, Klaus Wiehler, and Sven Kelling for helpful advice on the Trimontana spectrometers and technical assistance during implementation of the project. The survey of students was vetted by the Institutional review Board of Rice University (Protocol #FY2022-46), and performed by Christopher D. Barr at Rice University.

References

(1) Shepherd, B.; Bellamy, M. K. A Spreadsheet Exercise To Teach the Fourier Transform in FTIR Spectrometry. *J. Chem. Educ.* **2012**, *89* (5), 681–682. https://doi.org/10.1021/ed200547a.

(2) Dickson-Karn, N. M.; Orosz, S. Implementation of a Python Program to Simulate Sampling. *J. Chem. Educ.* **2021**, *98* (10), 3251–3257. https://doi.org/10.1021/acs.jchemed.1c00597.

(3) Maher, C.; Schazmann, B.; Gornushkin, I. B.; Rurack, K.; Gojani, A. B. Exploring an Application of Principal Component Analysis to Laser-Induced Breakdown Spectroscopy of Stainless-Steel Standard Samples as a Research Project. *J. Chem. Educ.* **2021**, *98* (10), 3237–3244. https://doi.org/10.1021/acs.jchemed.1c00563.

(4) Kovarik, M. L.; Clapis, J. R.; Romano-Pringle, K. A. Review of Student-Built Spectroscopy Instrumentation Projects. *J. Chem. Educ.* **2020**, *97* (8), 2185–2195. https://doi.org/10.1021/acs.jchemed.0c00404.

(5) Kuroki, N.; Mori, H. Comprehensive Physical Chemistry Learning Based on Blended Learning: A New Laboratory Course. *J. Chem. Educ.* **2021**, *98* (12), 3864–3870. https://doi.org/10.1021/ACS.JCHEMED.1C00666

(6) Wilson, M. V.; Wilson, E. Authentic Performance in the Instrumental Analysis Laboratory: Building a Visible Spectrophotometer Prototype. *J. Chem. Educ.* **2017**, *94* (1), 44–51. https://doi.org/10.1021/acs.jchemed.6b00515.

(7) Evans, J. S. O.; Evans, I. R. Structure Analysis from Powder Diffraction Data: Rietveld Refinement in Excel. *J. Chem. Educ.* **2021**, *98* (2), 495–505. https://doi.org/10.1021/acs.jchemed.0c01016.

(8) Keithley, R. B.; Sullivan, D. T.; Dodd, J. M.; Iyer, K. V.; Sarisky, C. A.; Johann, T. W. Learning about Fluorescence in Undergraduate Biochemistry: Enzyme Kinetics Using a Low-Cost, Student-Built Fluorescence Spectrometer. *J. Chem. Educ.* **2021**, *98* (12), 4054–4060. https://doi.org/10.1021/acs.jchemed.1c00912.

(9) Weiss, C. J.; Klose, A. Introducing Students to Scientific Computing in the Laboratory through Python and Jupyter Notebooks. *ACS Symp. Ser.* **2021**, *1387*, 57–67. https://doi.org/10.1021/BK-2021-1387.CH005.

(10) Sharma, A. K.; Thuermer, C.; Ruan, V. Learning Programming through Chemistry in a First-Year Scientific Computing Course. *ACS Symp. Ser.* **2021**, *1387*, 43–56. https://doi.org/10.1021/BK-2021-1387.CH004.

(11) Cohen, R.; Kelly, A. M. Community College Chemistry Coursetaking and STEM Academic Persistence. *J. Chem. Educ.* **2019**, *96* (1), 3–11. https://doi.org/10.1021/acs.jchemed.8b00586.

(12) Clinton, B. Opening College Doors to All Americans: Excerpts from Remarks at San Jacinto Community College. *J. Chem. Educ.* **1997**, *74* (12), 1392. https://doi.org/10.1021/ed074p1392.

(13) Weiss, C. J. Perspectives: Teaching Chemists to Code. *CEN Glob. Enterp.* **2017**, *95* (35), 30–31. https://doi.org/10.1021/cen-09535-scitech2.

(14) Chng, J. J. K.; Patuwo, M. Y. Building a Raspberry Pi Spectrophotometer for Undergraduate Chemistry Classes. *J. Chem. Educ.* **2021**, *98* (2), 682–688. https://doi.org/10.1021/acs.jchemed.0c00987.

(15) Albert, D. R. Constructing, Troubleshooting, and Using Absorption Colorimeters to Integrate Chemistry and Engineering. *J. Chem. Educ.* **2020**, *97* (4), 1048–1052. https://doi.org/10.1021/acs.jchemed.9b00548.

(16) Blum, A.; Hopcroft, J.; Kannan, R. *Foundations of Data Science*; Cambridge University Press, 2020.

(17) Wallum, A. *CAFF-Teaching Materials, https://github.com/alwallum/CAFF-Teaching-Materials*. https://github.com/alwallum/CAFF-Teaching-Materials.

(18) Busjahn, T.; Schulte, C. The Use of Code Reading in Teaching Programming. In *Proceedings of the 13th Koli Calling International Conference on Computing Education Research - Koli Calling '13*; ACM Press: Koli, Finland, 2013; pp 3–11. https://doi.org/10.1145/2526968.2526969.

(19) Adams-Mcnichol, A. L.; Shiell, R. C.; Ellis, D. A. Accurate, Photoresistor-Based, Student-Built Photometer and Its Application to the Forensic Analysis of Dyes. *J. Chem. Educ.* **2019**, *96* (6), 1143–1151. https://doi.org/10.1021/ACS.JCHEMED.8B00862.
(20) Freeman, S.; Eddy, S. L.; McDonough, M.; Smith, M. K.; Okoroafor, N.; Jordt, H.; Wenderoth, M. P. Active Learning Increases Student Performance in Science, Engineering, and Mathematics. *Proc. Natl. Acad. Sci. U. S. A.* **2014**, *111* (23), 8410–8415.
(21) Baugh, J. Assessment of Spreadsheet and Database Skills in the Undergraduate Student. *Inf. Syst. Educ. J.* **2004**, *2* (30), 1545-679X.
(22) Shrager, R. I.; Handler, R. W. Titration of Individual Components in a Mixture with Resolution of Difference Spectra, PKs, and Redox Transitions. *Anal. Chem.* **1982**, *52* (7), 1147–1152.
(23) Frans, S.D.; Harris, J. M. Least Squares Singular Value Decomposition for the Resolution of PK's and Spectra from Organic Acid/Base Mixtures. *Anal. Chem.* **1985**, *57* (8), 1718–1721.
(24) Shrager, R. I. Chemical Transitions Measured by Spectra and Resolved Using Singular Value Decomposition. *Chemom. Intell. Lab. Syst.* **1986**, *1* (1), 59–70.
(25) Grushow, A.; Reeves, M. S. Using Computational Methods to Teach Chemical Principles: Overview. *ACS Symp. Ser.* **2019**, *1312*, 1–10. https://doi.org/10.1021/BK-2019-1312.CH001.
(26) Metz, I. K.; Bennett, J. W.; Mason, S. E. Examining the Aufbau Principle and Ionization Energies: A Computational Chemistry Exercise for the Introductory Level. *J. Chem. Educ.* **2021**, *98* (12), 4017–4025. https://doi.org/10.1021/ACS.JCHEMED.1C00700.
(27) Esselman, B. J.; Hill, N. J. Integration of Computational Chemistry into the Undergraduate Organic Chemistry Laboratory Curriculum. *J. Chem. Educ.* **2016**, *93* (5), 932–936. https://doi.org/10.1021/ACS.JCHEMED.5B00815.
(28) Mcdonald, A. R. Teaching Programming across the Chemistry Curriculum: A Revolution or a Revival? *Teach. Program. Chem. Curric. Part 1 - Teach. Program. Chem. Curric. Revolut. Revival* **2021**. https://doi.org/10.1021/bk-2021-1387.ch001.
(29) Staveren, M. van. Integrating Python into a Physical Chemistry Lab. *J. Chem. Educ.* **2022**, *2022*. https://doi.org/10.1021/ACS.JCHEMED.2C00193.
(30) Virtanen, P.; Gommers, R.; Oliphant, T. E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; van der Walt, S. J.; Brett, M.; Wilson, J.; Millman, K. J. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nat. Methods* No. 17, 261–272.
(31) Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **2007**, *9* (3), 90–95. https://doi.org/10.1109/MCSE.2007.55.
(32) McKinney, W. Data Structures for Statistical Computing in Python. In *Proceedings of the 9th Python in Science Conference*; Austin, TX, 2010; Vol. 445, pp 51–56.
(33) Harris, C. R.; Millman, K. J.; van der Walt, S. J.; Gommers, R.; Virtanen, P.; Cournapeau, D.; Wieser, E.; Taylor, J.; Berg, S.; Smith, N. J.; Kern, R.; Picus, M.; Hoyer, S.; van Kerkwijk, M. H.; Brett, M.; Haldane, A.; del Río, J. F.; Wiebe, M.; Peterson, P.; Gérard-Marchant, P.; Sheppard, K.; Reddy, T.; Weckesser, W.; Abbasi, H.; Gohlke, C.; Oliphant, T. E. Array Programming with NumPy. *Nature* **2020**, *585* (7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2.
(34) Wang, J. J.; Rodríguez Núñez, J. R.; Maxwell, E. J. Build Your Own Photometer: A Guided-Inquiry Experiment To Introduce Analytical Instrumentation. *J. Chem. Educ.* **2016**, *93* (1), 166–171.
(35) National Academies of Sciences, E. and M. Data Science: Opportunities to Transform Chemical Sciences and Engineering: Proceedings of a Workshop—in Brief. *Data Sci.* **2018**. https://doi.org/10.17226/25191.