JOURNAL OF CHEMICAL EDUCATION

pubs.acs.org/jchemeduc Article

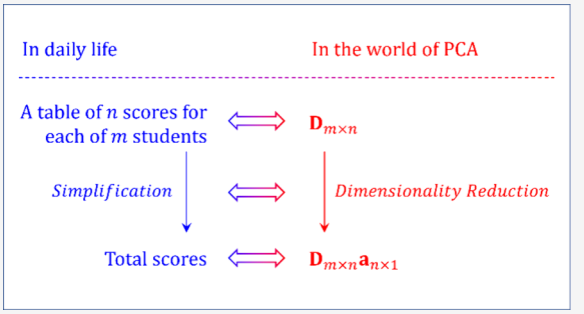
Teaching Principal Component Analysis in the Course of Analytical Chemistry: A Q&A Based Heuristic Approach

Limin Shao*

Cite This: J. Chem. Educ. 2025, 102, 155−163 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Principal component analysis (PCA) is a powerful tool to process matrix-type data that is more and more common in analytical chemistry, so the necessity to teach PCA becomes more and more evident. Students can easily implement PCA with modern software, but they need deep understanding of the theory to avoid misuses and misinterpretations. However, the theory of PCA is fairly difficult. In order to ensure effective teaching, we have designed a series of questions and answers (Q&As) that begin with a daily life example and end at the core logic of PCA. These Q&As form a heuristic chain for students to follow, so that they can gradually approach and eventually comprehend the theory of PCA. Once students are equipped with such a comprehension, it is easy and natural to further teach PCA properties and applications. For that purpose, we prepared four typical examples as well as in-depth discussions. Positive feedbacks from students confirm the effectiveness of this approach.



KEYWORDS: *Analytical Chemistry, Principal Component Analysis, Teaching PCA*

1. INTRODUCTION

Principal component analysis (PCA) is a powerful tool to process complex matrix-type data. It has been widely recognized in various fields. A search on the Web of Science (topic: principal component analysis; database: Core Collection) shows 92,925 publications from 2013 to 2023, and the top five categories are environmental sciences (11%), engineering electrical electronics (10%), food science technology (7%), computer science artificial intelligence (6%), and analytical chemistry (5%).

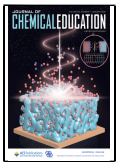
Modern instruments usually measure too many features for a finite number of samples. The phenomenon that measured features outnumber samples is referred to as *curse of dimensionality*.[1] The curse of dimensionality is more and more common in analytical chemistry because of instrumental advances, which necessitate methods of dimensionality reduction such as PCA.

In analytical chemistry, PCA is applied to chromatography,[2] molecular spectrometry,[3,4] nuclear magnetic resonance,[5] mass spectrometry,[6,7] imaging,[8−11] and bioinformatics.[12] Moreover, it is the base of factor analysis methods.[13] Such wide applications encourage teachers to explore teaching PCA, even though it is not included in the traditional course of analytical chemistry.[14−24] Most explorations use various examples to demonstrate the applications of PCA instead of focusing on teaching the theory.

Since 2014 we started teaching PCA in a one semester course entitled "Chemometrics" for fourth-year undergraduate and first-year graduate students of chemistry. The number of students was about 15 to 20 in the beginning and now reaches 70 with more graduate students. Students enrolling in the course are required to have basic knowledge of calculus and linear algebra, such as partial derivatives and matrix operations. In 2023, the Ministry of Education of China initiated a nationwide project to reform higher education named the "101 Plan". The "101 Plan" of Chemistry includes PCA in the course of Analytical Chemistry. Analytical chemistry is a compulsory course in most Chinese universities for first- or second-year students of chemistry who already have the prerequisite knowledge.

Although we started teaching PCA a long time ago, it was not until recently that we came up with an effective approach. In this approach, students are given a question and then guided by the teacher to form the answer that leads to another question. These Q&As were elaborately designed in order to form a heuristic chain. By following the chain, students gradually get insight into PCA, from easy aspects to more difficult aspects and from surface understanding to deep understanding. This approach is a result of our explorative attempts.

Received: July 3, 2024
Revised: December 4, 2024
Accepted: December 5, 2024
Published: December 17, 2024



Our first attempt to teach PCA was naturally application-focused. It was easy for the teacher because of many application examples accumulated in the long history of PCA. It was also easy for the students because user-friendly software implements PCA with just a few mouse clicks. We observed noticeable progress in the students who quickly learned how to perform PCA, visualize results, and interpret conclusions. Such a success is partly because we deliberately allocated more time to applications rather than to the theory of PCA. A brief introduction to the theory did save some time but resulted in shallow understanding and consequent disadvantages. For instance, students learned the well-known notion of dimensionality reduction by PCA but had difficulty in grasping the core; they could recite the well-known sentence of "*the first principal component accounts for the most variance*" but did not understand the mechanism or the role of variance. In this way, students learned PCA as a black-box. So, if at a later time they do not have chances to use PCA, they will likely forget how to implement it, and what they had learned would not change into knowledge to strengthen their scientific thinking.

Our second attempt to teach PCA was an upgrade from the first one, with emphasis on the theory. We prepared a complete scheme to cover important aspects of the theory, but unfortunately, the result was worse. Students told us that they were quickly lost in expressions of linear algebra; even after they managed to understand those, they found difficulties in grasping the underlying logic. Students doubted the merits of pondering on the theory; after all, they could easily perform PCA without much understanding of it. Contrary to what we had planned, a detailed lecture on the theory was no more effective than a brief introduction.

The two aforementioned attempts created a dilemma in teaching PCA; i.e., the theory should be taught in detail for deep understanding, while linear algebra is an overwhelming obstacle. After a while we realized that the theory could be effectively delivered if the fundamental logic is sufficiently explained. Such an explanation could be prepared in plain texts. Plain texts are not as precise as linear algebra expressions, but they are readily understandable, particularly with analogies that students are familiar with in daily life. Once students understand the fundamental logic of PCA, they would not resist linear algebra expressions in the theory, because (1) the expressions represent what they have already understood and (2) the expressions are precise and concise. After students have sufficient knowledge of the theory of PCA, further teaching its properties and applications becomes easy. This method is consistent with the teaching philosophy of Richard P. Feynman who proposed avoiding definitions in the first lesson despite their importance, and his words are "*That is just an example of the difference between definitions (which are necessary) and science. The only objection in this particular case was that it was the first lesson.*"[25]

2. TEACHING THE FUNDAMENTAL LOGIC OF PCA THROUGH A Q&A BASED HEURISTIC CHAIN

PCA was invented by Karl Pearson in 1901[26] and independently developed and named in 1933 by Harold Hotelling.[27] Since then, PCA has been gradually and widely applied in various fields.

This brief introduction was meant not only to describe the origin of PCA but also to lead to the first question of the following Q&A chain naturally. The Q&A chain could be conducted between the teacher and students or solely by the teacher.

- **Q:** Why is PCA taught in the course of analytical chemistry long after it was invented?
- **A:** Experimental data in analytical chemistry used to be fairly simple, and did not need powerful tools like PCA to process. The situation changed when advanced instrumentation, such as GC-MS, appeared and generated matrix-type data. Matrix is the target of PCA.
- **Q:** Matrix sounds quite abstract to understand.
- **A:** It is, but we have a good helper. Suppose a score table of m students who each have n scores, it is an m-by-n matrix, and let's denote it as $\mathbf{D}_{m\times n}$.
- **Q:** $\mathbf{D}_{m\times n}$ seems abstract, but easy to understand when associated with a real-world example. It is also easy to process a score table, e.g., to calculate total scores. Does such a calculation have any matrix-related meaning?
- **A:** Calculating total scores from score table $\mathbf{D}_{m\times n}$ can be concisely expressed with matrix multiplication $\mathbf{D}_{m\times n}\mathbf{a}_{n\times 1}$, where $\mathbf{a}_{n\times 1}$ is an n-by-1 vector of all ones. Interestingly, vector $\mathbf{a}_{n\times 1}$ not only yields total scores, but also simplifies $\mathbf{D}_{m\times n}$, given the fact that $\mathbf{D}_{m\times n}$ has $m \times n$ values (the original scores), whereas $\mathbf{D}_{m\times n}\mathbf{a}_{n\times 1}$ has m values (the total scores).
- **Q:** We have not noticed that calculating total scores also means simplification, but it does make sense. When $\mathbf{D}_{m\times n}$ is simplified into $\mathbf{D}_{m\times n}\mathbf{a}_{n\times 1}$, the number of values is reduced from $m \times n$ to m. So, is there any consequent information loss?
- **A:** You are right, there is information loss.
- **Q:** It has never occurred to us that calculating total scores incurs information loss. Can we prevent such loss?
- **A:** Yes, we can. Do not calculate total scores, just use the original score table, then there will be no information loss.
- **Q:** Is it a joke? How are we supposed to use the score table directly?
- **A:** We cannot directly use the score table because of its complexity, so we calculate total scores as a way to simplify it. This natural operation reflects the core of PCA that is called *dimensionality reduction*. Dimensionality reduction essentially simplifies a data matrix that is too complex to process directly.
- **Q:** Now we understand that simplification, or dimensionality reduction for that matter, is a necessary step to process a complex data matrix. But the consequent information loss still concerns us. How does PCA handle this problem?
- **A:** Although information loss is inevitable, PCA could keep as much information as possible. This feature also earns its name, which suggests the ability to produce the principal component of the original matrix.
- **Q:** While we are wondering how PCA could keep the most information, we are also eager to know how you measure the amount of information.
- **A:** Let's look at an example (draw a horizontal line and a random curve on the blackboard). We would say that the curve has "more" information than the straight line; moreover, the more the curve varies, the more information it has.
- **Q:** It is reasonable to link the amount of information with variation. Then, how is variation evaluated?
- **A:** Variation is usually evaluated with a statistical concept named variance. Variance is defined as $\frac{\sum_i^n (x_i - \overline{x})^2}{n-1}$, where x_i represents one of the n values, and $\overline{x}$ is the average.

https://doi.org/10.1021/acs.jchemed.4c00818
J. Chem. Educ. 2025, 102, 155−163

- **Q:** We seem to understand the logic of PCA now. PCA simplifies data matrix $\mathbf{D}_{m\times n}$ that is too complex to process directly. The simplification is implemented by multiplying it with vector $\mathbf{a}_{n\times 1}$, and the result $\mathbf{D}_{m\times n}\mathbf{a}_{n\times 1}$ somehow has the largest variance, which ensures the retention of the most information. Is that right?
- **A:** You are totally right, and you already know enough to understand some popular statements about PCA, such as "$\mathbf{D}_{m\times n}\mathbf{a}_{n\times 1}$ *contains as much information as possible*", "$\mathbf{D}_{m\times n}\mathbf{a}_{n\times 1}$ *accounts for the most variance*".
- **Q:** Vector $\mathbf{a}_{n\times 1}$ plays a key role to keep the most information, how can we obtain it?
- **A:** Vector $\mathbf{a}_{n\times 1}$ is obtained through a rigorous mathematic procedure, which is explained in section #3 of the Supporting Information. At this moment, just keep in mind that there exists such a vector, and it ensures that $\mathbf{D}_{m\times n}\mathbf{a}_{n\times 1}$ keeps the most information of $\mathbf{D}_{m\times n}$.
- **Q:** How about the information that is not kept in $\mathbf{D}_{m\times n}\mathbf{a}_{n\times 1}$?
- **A:** We find another vector, say $\mathbf{b}_{n\times 1}$, so that $\mathbf{D}_{m\times n}\mathbf{b}_{n\times 1}$ keeps the most of the rest of the information. We also say that $\mathbf{D}_{m\times n}\mathbf{b}_{n\times 1}$ keeps the most information of $\mathbf{D}_{m\times n}$, and it is irrelevant to $\mathbf{D}_{m\times n}\mathbf{a}_{n\times 1}$. $\mathbf{D}_{m\times n}\mathbf{a}_{n\times 1}$ is called the *first principal component* (PC1), and $\mathbf{D}_{m\times n}\mathbf{b}_{n\times 1}$ is called the *second principal component* (PC2). Interestingly, PCs are also called score vectors, or scores. The percentage of the information kept by a PC is the ratio of its variance over the total variance of all PCs. The percentage can be conveniently calculated with $\frac{\lambda_i}{\Sigma\lambda_j}$, where λ_i is the eigenvalue corresponding to the ith PC. Explanations of eigenvalues and eigenvectors were presented in section #3 of the Supporting Information.
- **Q:** Is there a third PC, i.e., PC3? How many PCs are there?
- **A:** The number of PCs is mathematically determined, and explained in section #3 of the Supporting Information. At this moment, just keep in mind that there are sufficient PCs, and they together contain all the original information.
- **Q:** Then, if we keep all PCs, is there still information loss?
- **A:** If so, there will be no information loss, but there will be no simplification either.
- **Q:** Is information loss the price we pay to simplify a complex data matrix?
- **A:** Exactly. Information loss is inevitable, but PCA can minimize the loss. In other words, if you want to simplify matrix $\mathbf{D}_{m\times n}$ into a single column, PC1 is your best choice; if you want to simplify $\mathbf{D}_{m\times n}$ into two columns, PC1 together with PC2 is your best choice, and so on.
- **Q:** All right, the fundamental logic of PCA seems clear, anything else?
- **A:** We are done. Please be noted that this conversation is intended to pave the road to understanding the rigorous mathematics of the theory, not to replace it. Everyone is encouraged to study the theory sheet. Mathematics and its symbolic system are crucial and indispensable to science.
- **Q:** Wait, since PC1 contains the most information, why shouldn't we select PCA over calculating total scores to process a score table?
- **A:** PC1 undoubtedly has the most information, as long as the amount of information is measured with variance. Variance is a good measure, but not an exclusive one. To simplify a score table, calculating total scores is fine and very convenient.
- **Q:** Do you imply that PCA is difficult to compute?
- **A:** With advanced software like Matlab or Octave, it only takes a few commands to perform PCA, which you can find in the theory sheet. What really matters is interpretations of the results, and that requires a deep understanding of the theory.

The theory sheet of PCA can be found in section #3 of the Supporting Information, and it is given to students after the Q&A script is finished. In both the above explanation and the theory of PCA, no assumptions were made. However, if PCA is used for the purpose of statistical inference, normal distribution of data is desired.[28]

For the Q&A script to be used in a live instructional session, it would have to be controlled by the instructor, so that students are in the right track planned toward understanding the theory of PCA. In our practice, the lecture of this part takes 45 min, the length of a class in most Chinese universities.

3. ILLUSTRATING PROPERTIES OF PCA WITH EXAMPLES

Once students understand the fundamental logic of PCA, they are ready to learn properties of PCA. With such knowledge, they could correctly interpret the results and more importantly avoid misusing PCA. Four aspects of the knowledge are particularly elaborated in our class with selected examples. The four aspects were chosen for the sake of balancing the theory and applications of PCA; students should grasp such knowledge for the proper applications of PCA.

After class, all data of the examples are given to students to practice with. The students are asked to reproduce the results that the instructor presents during the lecture and encouraged to use the codes in section #3 of the Supporting Information.

3.1. Mean-Centering or Not?

Mean-centering a matrix subtracts the mean of each column from all elements of the column. So, for a mean-centered matrix, each column has a zero mean.

Mean-centering is the first step of PCA. In some cases, however, this step is omitted either unintentionally or deliberately, and the results might not be significantly different. This phenomenon and the role of mean-centering can be explained by the theory of PCA.

As explained in the first section, the measure of information in PCA is variance expressed as $\frac{\sum_i^n (x_i-\bar{x})^2}{n-1}$. This expression implies mean-centering, and without mean-centering, it becomes $\frac{\sum_i^n x_i^2}{n-1}$, referred to as mean squared Euclidean norm (MSEN). Therefore, if the step of mean-centering is omitted from PCA, the measure of information changes from variance to MSEN.

When variance is the measure of information, a signal with large fluctuations is considered information-rich regardless of intensities. When MSEN is the measure of information, a signal with large intensities is considered information-rich, regardless of fluctuations.

For practical signals, fluctuation often relates to intensity; therefore, high variance also means high MSEN, and vice versa. For such signals, the results of PCA with or without mean-centering are fairly similar.

There do exist signals with high intensities and low fluctuations, e.g., baselines. If such signals are present in the

https://doi.org/10.1021/acs.jchemed.4c00818
J. Chem. Educ. 2025, 102, 155−163

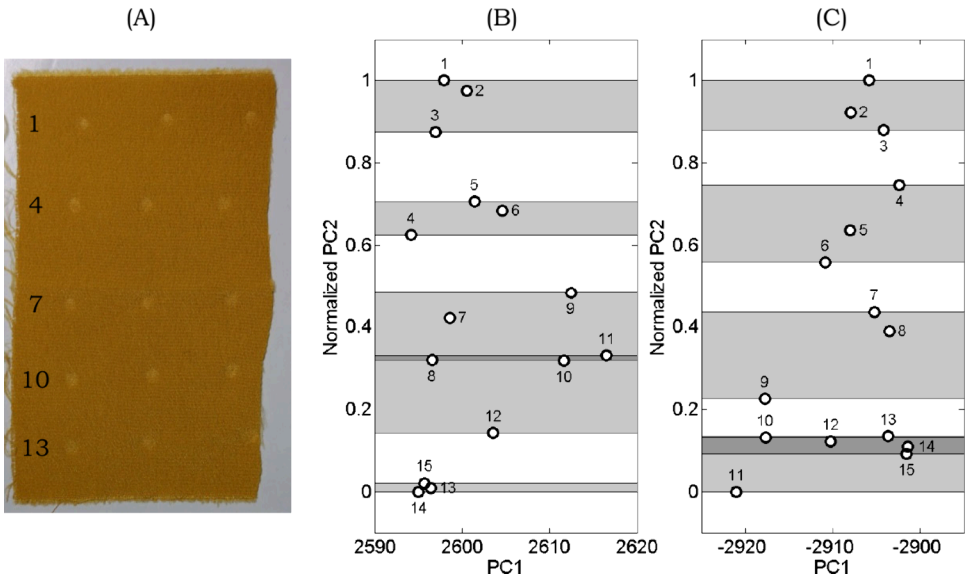


Figure 1. (A) Photograph of a piece of silk fabric, on which 15 IR spectra were measured; the first sampling mark in each of the 5 horizontal lines is numbered for reference. (B and C) Scatter plots from PCA with and without mean-centering the data matrix, respectively. Dots in (B) or (C) correspond to sampling marks in (A). Each gray stripe in (B) or (C) is formed by the three dots of the same horizontal line in (A); darker parts indicate overlapping.

matrix, they are probably lost in the results of PCA with mean-centering but kept in the results of PCA without mean-centering.[29]

Figure 1A shows a piece of dyed silk fabric, and the marks are the 15 locations at which infrared spectra (IR) were measured. Each of the 15 spectra has 831 data points within the wavenumber range from 1800 to 1000 cm^{-1}. The spectra were arranged rowwise to form a 15-by-831 data matrix. The matrix was processed with PCA to reveal possible relationships between the IR spectra and their sampling locations. Data and Matlab codes were provided in the Supporting Information.

Figures 1B and 1C show the results of PCA with and without mean-centering, respectively. There are five stripes in Figure 1B, each of which is formed by the three dots of the same horizontal line in Figure 1A. There are only four stripes in Figure 1C; the stripe formed by dots #10, #11, and #12 completely overlaps with the stripe formed by dots #13, #14, and #15. PCA with and without mean-centering measures information by variance and MSEN, respectively, so the characteristics in Figures 1B and 1C are controlled by large variations and large numerical values of the variables. This led us to inspect the spectra, revealing large variations and large values between 1674 and 1481 cm^{-1}. We deleted these data from the matrix, performed PCA, and found that the stripes along the PC2 axis were better separated, especially by PCA without mean-centering, as shown in Figure S1. Therefore, spectral data outside the ranges of 1674 and 1481 cm^{-1} are more helpful to relate PC2 to the sampling locations of the spectra. Data and Matlab codes for Figure S1 were provided in the Supporting Information.

In summary, mean-centering is an integral part of the standard procedure of PCA, yet PCA without mean-centering should not be considered wrong; instead, it is another way of simplifying a matrix to explore information. Brereton has extensively studied several cases, in which results from mean-centered matrices differ from those from non-mean-centered matrices.[30] The role of mean-centering is also discussed in quantitative analysis.[28,31]

In some references or books on PCA, mean-centering is regarded as a preprocessing step. Another common preprocessing method is scaling, which divides variables (columns of the data matrix) by the corresponding standard deviations. Scaling makes variables in different units more comparable.[28] In the sense of preprocessing, there are several methods, such as smoothing, baseline removal, derivative transformation, etc. It is noteworthy that mean-centering is an integral part of PCA, whereas preprocessing methods are not, and their effects are often case-dependent.

In our teaching practice, this part takes about 20 min with emphasis on the meaning and the operation of mean-centering and the effect on the loss of certain information after PCA.

3.2. The First Pattern of Simplification by PCA

When we focus on principal components instead of the original matrix, we follow the first pattern of simplification by PCA. In this pattern, the original matrix is simplified into a few PCs for subsequent analysis, and the most common way is plotting PC2 versus PC1.

Such a convenient 2D plot comes at the cost of information loss, which might result in significant biases in any conclusions drawn from the plot. Therefore, such a plot usually includes the percentage of the information kept in PC1 (or PC2), that is, the ratio of the variance of PC1 (or PC2) to the total variance of the original matrix.

35 samples of 3 components were simulated; concentrations of the 3 components are listed in Table S1 of the Supporting Information. The table provides a little information about the 35 samples due to its complexity. So, the matrix was simplified by PCA, and PC2 was plotted against PC1 in Figure 2. Data and Matlab codes were provided in the Supporting Information.

Figure 2 clearly shows 7 distinct groups, which guides us to rearrange the original data of Table S1 accordingly. The rearranged data in Table S2 reveal obvious similarity in each group, so we conclude that the original samples can be classified into 7 groups. Without PCA, such a classification would be fairly difficult to obtain directly from the original 35-by-3 matrix.

According to the axis labels in Figure 2, PC1 and PC2 together contain 77.8% of the original information. The addition of the

https://doi.org/10.1021/acs.jchemed.4c00818
J. Chem. Educ. 2025, 102, 155−163

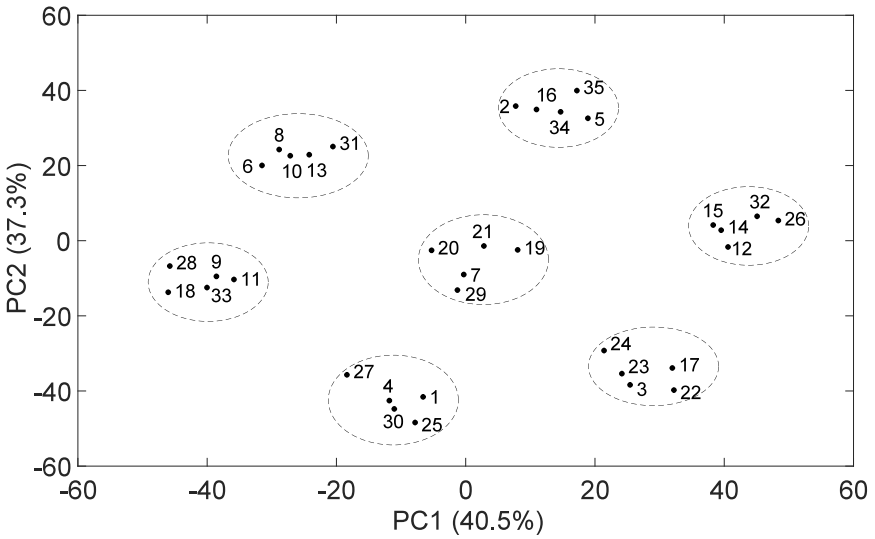


Figure 2. Scatter plot of PC2 versus PC1. Each sample is represented by a dot, and next to it is the sample number. Aggregated dots are encircled for clarity. The two parenthesized values in axis labels are the percentages of information kept in PC1 and PC2, respectively.

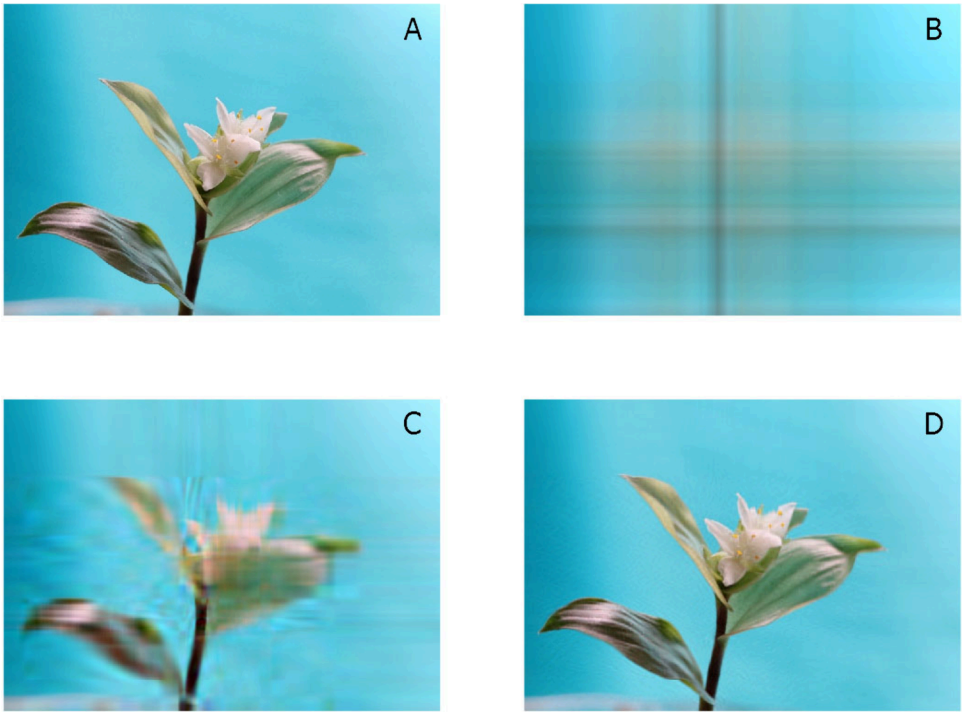


Figure 3. (A) The original picture; (B, C, and D) pictures constructed with 1, 11, and 57 PCs, respectively. PCs were obtained by performing PCA with the RGB matrices of the original picture. All 4 pictures have the same resolution of 700 × 500.

two percentages should not be taken for granted; it is allowed because PCs are irrelevant.

This example is meant for illustration. The original matrix can be visualized in a 3D plot without information loss, but it is less clear than the 2D plot in Figure 2. If a matrix has more than 3 columns, which is fairly common in practice, it cannot be visualized directly, whereas visualizations by PCA are feasible and convenient.

This example shows the applicability of PCA to clustering, another method of which is *k*-means. Both PCA and *k*-means are unsupervised, as opposed to supervised methods, such as *k*-nearest neighbors (*k*-NN), which are used for classification. Conceptually, *supervised* and *unsupervised* refer to whether samples are labeled or unlabeled, with concentrations of samples, for instance. In other words, supervised methods need extra information to yield desired results, which is not the case for unsupervised methods like PCA.

In our teaching practice, this part takes about 20 min with emphasis on the logic of the simplification by PCA, the calculation of information loss, and the presentation of the results.

3.3. The Second Pattern of Simplification by PCA

The second pattern of simplification by PCA includes 3 steps.

1. Perform PCA to the original matrix $\mathbf{D}_{m\times n}$.

https://doi.org/10.1021/acs.jchemed.4c00818
J. Chem. Educ. 2025, 102, 155−163

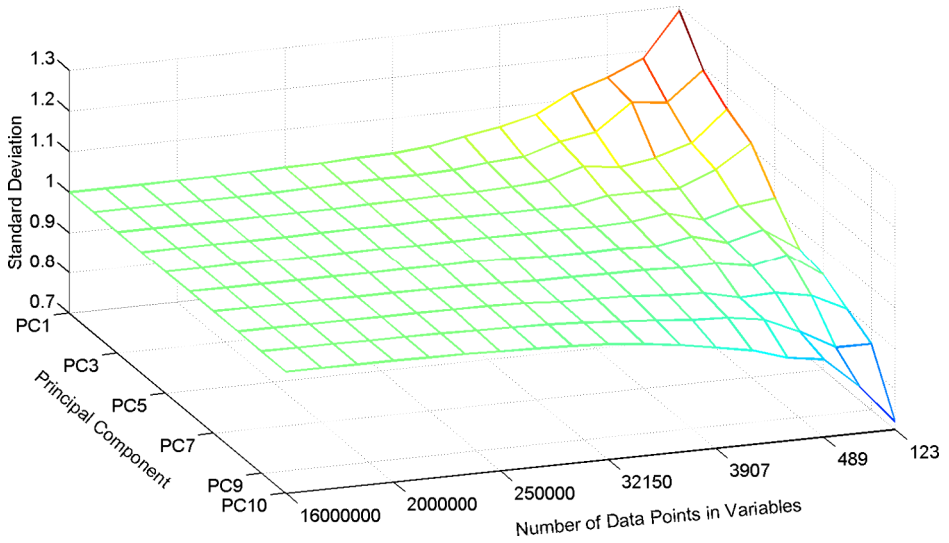


Figure 4. Standard deviations of 10 principal components from 18 matrices. Each matrix comprises 10 random variables of standard normal distribution (zero mean and unit variance). Sizes of the matrices are x by 10, where x starts from 16,000,000, is halved each time, and ends at 123.

2. Arrange the first p PCs in the descending order of variance to form a matrix $\mathbf{U}_{m\times p}$ in a column-wise manner. Do the same to the corresponding p eigenvectors to form another matrix $\mathbf{V}_{n\times p}$.
3. Use $\mathbf{D}^{\#}_{m\times n} = \mathbf{U}_{m\times p}(\mathbf{V}_{n\times p})^{t}$ to construct matrix $\mathbf{D}^{\#}_{m\times n}$, where superscript t denotes the transpose. Matrix $\mathbf{D}^{\#}_{m\times n}$ is used instead of the original one $\mathbf{D}_{m\times n}$ for subsequent analysis.

If $\mathbf{D}^{\#}_{m\times n}$ is constructed with all PCs, then it is identical to the original one $\mathbf{D}_{m\times n}$; if not, some information is lost. The lost information, according to the theory of PCA, has a low variance. Therefore, this pattern of simplification could remove some low-variance information that in experimental data usually means noise. Obviously there should be sufficient PCs to construct $\mathbf{D}^{\#}_{m\times n}$ so that it contain as much needed information as possible. There are several criteria that determine the number of sufficient PCs.[13] One simple criterion is called Kaiser's criterion, also known as the scree graph. In Kaiser's criterion, the original matrix is standardized to have zero means and unit variances, and PCs with eigenvalues greater than one are determined to be sufficient and kept.

This pattern of simplification was illustrated by the processing of a picture. The data of a picture are three matrices that contain red, green, and blue information on all pixels, so it is an ideal material to demonstrate matrix manipulations and result visualizations.

The R, G, and B matrices in Figure 3A were processed by PCA, respectively. For each of the three matrices, it was found that the first PC contains more than 93% of the information, the first 11 PCs contain more than 99% of the information, and the first 57 PCs contain more than 99.9% of the information. Pictures constructed with 1, 11, and 57 PCs are shown in Figures 3B, 3C, and 3D, respectively.

Figure 3B is far from the original picture despite the fact that the first PC contains 93% of the information. When 11 PCs were used to construct the picture, major details were restored, as shown in Figure 3C, but distortions were visible. When 57 PCs were used, the constructed picture is visually the same as the original one, as shown in Figure 3D.

In Figure 3, the original picture and the constructed ones all have the same resolution of 700 × 500 and the same number of data (700 × 500 × 3 = 1,050,000), which does not seem to imply simplification. In fact, all three constructed pictures are from simplified data. For example, Figure 3D was constructed with 57 PCs (each has 500 data) and 57 eigenvectors (each has 700 data), so the total number of data is $[(57\times 500) + (57\times 700)]\times 3 = 205{,}200$, which means it only took ∼20% of the original data to construct a picture close enough to the original one.

In our teaching practice, this part takes about 20 min with emphasis on the logic of the simplification by PCA, the difference between this pattern and the first one explained in section 3.2. In the first pattern, both low-variance information and the number of data points are reduced by PCA. In the second pattern, only low-variance information is reduced, and the constructed matrix has the same size as the original one.

3.4. Noise Reduction by PCA

In experimental data, noise usually has lower variance than the interest signals, so it is considered less principal by PCA (with mean-centering!) and less included in fore-end PCs, e.g., PC1, or PC2, etc. This phenomenon is noise reduction. Mathematical analysis can be found in refs 32 and 33.

Noise reduction by PCA seems natural, but the actual situation is much more complicated. It turns out that noise reduction by PCA is governed by the Law of Large Numbers; when there is not a large number of data points, noise after PCA might even be increased. Theoretical analysis presented in section #4 of the Supporting Information is for students to study after class, while simulated data are used for demonstration in the class.

Matrices of different sizes were prepared with random values from standard normal distribution. The sizes are set to be 16,000,000 by 10, 8,000,000 by 10, 4,000,000 by 10, ..., and 132 by 10, respectively, and there are 18 matrices in total. These matrices contain pure noise, the noise level of which is unanimously one, and the only difference is the number of data points.

The 18 matrices were processed by PCA. PCA with and without mean-centering yielded the same results since data were from standard normal distribution with zero mean. The standard deviations of the first 10 PCs were calculated to measure noise levels, shown as a 3D plot in Figure 4. The 3D plot shows clearly that, when there are a large number of data points, e.g., 16,000,000 and 8,000,000, the noise levels of the 10 PCs are

https://doi.org/10.1021/acs.jchemed.4c00818
J. Chem. Educ. 2025, 102, 155−163

almost the same and equal to 1. This result is consistent with eq 7 of the Supporting Information. Data and Matlab codes were provided in the Supporting Information.

When the number of data points decreases, the noise level of PC1 is larger than 1, as shown in Figure 4. Specifically, for matrices with 62,500, 3,907, and 977 rows, the noise level of PC1 increases by 1%, 5%, and 10%; when the number of data points is 123, the noise level of PC1 reaches 1.3, 30% higher than that of the original noise level. The increase in the noise level was also observed for PC2. For rear-end PCs, such as PC9 or PC10, the noise level drops with a decrease in data points. The total variance of all PCs equals 10, so higher variance for fore-end PCs means lower variance for rear-end PCs.

In summary, the statistical independence of noise requires a large number of data points in the matrix. If so, noise is distributed homogeneously among all PCs; if not, noise is treated as nonrandom information and included more in fore-end PCs to maximize variance (due to the nature of PCA). For the latter case, PCA actually increases the noise level of fore-end PCs.

In analytical chemistry, the number of rows of a matrix equals the number of chemical samples. Therefore, the well-known advice in analytical chemistry of preparing as many samples as possible has another benefit in PCA, which is to prevent PCA from including more noise in fore-end PCs.

It should be noted that the above conclusion applies only to the first pattern of simplification by PCA explained in section 3.2. In the second pattern of simplification explained in section 3.3, noise reduction is always observed in matrices constructed from fore-end PCs.[34,35]

This section closely relates to the first and the second pattern of simplification by PCA, and thus serves as a valuable material for better and deeper understanding of PCA. This section is important but mathematically challenging, so it is for students with a strong mathematical background, not compulsory for everyone.

In our teaching practice, this part takes about 15 min with emphasis on the logic of the noise reduction by PCA, the reason that noise reduction is generally not expected in the first pattern of simplification by PCA.

4. EXERCISE FOR AND FEEDBACK FROM STUDENTS

Eight questions were handed out to students for them to practice with, which cover the knowledge of PCA described in the manuscript. These questions and suggested answers were provided in section #5 of the Supporting Information.

In 2020 when the method was not adopted, 72 students enrolled in the course and evaluated with the test provided in section #7 of the Supporting Information. The average score was 82.1 with a standard deviation of 5.2. In 2024 when the method was fully adopted, 99 students enrolled in the course and evaluated with the same test for comparison. The average score was 85.5 with a standard deviation of 5.5. Shapiro−Wilk tests showed that neither of the two score sets was normally distributed, so the Mann−Whitney U-test, a nonparametric approach, was used to compare the two score sets instead of *t*-test. Results showed that the median score in 2024 was statistically greater than that in 2020 (P-value < 0.001). Students' scores and Matlab codes to perform Mann−Whitney U-test were provided in the Supporting Information. Moreover, a survey was conducted to evaluate the method's efficiency.

In the spring semester of 2024, a survey was carried out with 54 students to evaluate the effectiveness of the approach. 45 students (83%) considered it very helpful, 9 students (17%) found it helpful, and none reported it as unhelpful. The consensus among most remarks is that the approach makes abstract PCA concepts, like matrix and dimensionality reduction, more concrete with a score table and the calculation of total scores that are both straightforward and interesting. Here is an example of the remarks:

> *"I think this approach is very helpful. Firstly, with the help of a score table, we can directly understand why dimensionality reduction is needed and how it is implemented. Secondly, with the help of a score table, we can clearly observe the operations and mathematics involved in each step of PCA. Thirdly, we are familiar with score tables and their processing, which makes learning PCA less difficult and boring."*

Interestingly, a student interpreted a score table from the perspective of PCA, which definitely results from a good grasp of the theory:

> *"I think this approach is very helpful for several reasons. Students are familiar with a score table, so they could intuitively understand primary concepts of PCA; for example, each subject is one variable, and scores of a student is a sample of all variables. In a score table, some subjects may be correlated, e.g., scores of Mathematics and Physics, and PCA yields a combination of the scores that keeps the most discrepancies."*

5. CONCLUSIONS

Principal component analysis (PCA) is commonly used nowadays in analytical chemistry as a powerful tool to process matrix-type data, so it has been taught in the course of analytical chemistry at our university. Understanding the theory of PCA is necessary to avoid misuses and misinterpretations, but it is fairly difficult. In order to ensure effective teaching, we have designed a series of questions and answers that begin with a daily life example and end at the core logic of PCA. These Q&As form a heuristic chain for students to follow, so that they can gradually approach and eventually comprehend the fundamental logic of PCA. Then we teach some important properties, which becomes less difficult for students after they have adequate knowledge of the theory.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available at https://pubs.acs.org/doi/10.1021/acs.jchemed.4c00818.

Simulated concentrations (arbitrary unit) of 3 components in each of 35 samples, those samples rearranged based on the group information from PCA, the theory of PCA, the relationship between the number of matrix rows and the noise level of PCs, the questions for students to practice with, Figure S1, and tests to evaluate students in 2020 and 2024 (PDF, DOCX)

Data and Matlab codes of Figure 1 (ZIP)

Data and Matlab codes of Figure 2 (ZIP)

Data and Matlab codes of Figure 4 (ZIP)

Data and Matlab codes of Figure S1 (ZIP)

Student scores in 2020 and 2024 and Matlab codes for Mann−Whitney U-test (ZIP)

https://doi.org/10.1021/acs.jchemed.4c00818
J. Chem. Educ. 2025, 102, 155−163

AUTHOR INFORMATION

Corresponding Author

Limin Shao − *Department of Chemistry, University of Science and Technology of China, Hefei, Anhui 230026, P. R. China;* orcid.org/0000-0002-7416-3692; Email: lshao@ustc.edu.cn

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jchemed.4c00818

Notes

The author declares no competing financial interest.

ACKNOWLEDGMENTS

This work was funded by the Teaching Team Project of Anhui Province (Grant No. 2021jxtd322) and "101 Plan" of Anhui Province (Grant No. 2023ylyjh001). The author would like to thank Dr. Wanping Wang for the assistance of associated literature. The author would also like to thank the editors and the reviewers for their valuable comments to improve the manuscript.

REFERENCES

(1) Curse of dimensionality on Wikipedia. https://en.wikipedia.org/wiki/Curse_of_dimensionality (accessed November 16, 2024).

(2) Nanayakkara, Y. S.; Woods, R. M.; Breitbach, Z. S.; Handa, S.; Slaughter, L. M.; Armstrong, D. W. Enantiomeric Separation of Isochromene Derivatives by High-Performance Liquid Chromatography Using Cyclodextrin Based Stationary Phases and Principal Component Analysis of the Separation Data. *J. Chromatogr. A* **2013**, *1305*, 94−101.

(3) Lehtinen, J.; Hirschmann, C. B.; Keiski, R. L.; Kuusela, T. Human Hair in the Identification of Cocaine Abuse with Cantilever-Enhanced Photoacoustic Spectroscopy and Principal Component Analysis. *Appl. Spectrosc.* **2013**, *67* (8), 846−850.

(4) Wang, W.; Shao, L.; Yuan, B.; Zhang, X.; Liu, M. Determining the number of chemical species in nuclear magnetic resonance data matrix by taking advantage of collinearity and noise. *Anal. Chim. Acta* **2018**, *1022*, 20−27.

(5) Gu, H.; Pan, Z.; Xi, B.; Asiago, V.; Musselman, B.; Raftery, D. Principal Component Directed Partial Least Squares Analysis for Combining Nuclear Magnetic Resonance and Mass Spectrometry Data in Metabolomics: Application to the Detection of Breast Cancer. *Anal. Chim. Acta* **2011**, *686*, 57−63.

(6) Corilo, Y. E.; Podgorski, D. C.; McKenna, A. M.; Lemkau, K. L.; Reddy, C. M.; Marshall, A. G.; Rodgers, R. P. Oil Spill Source Identification by Principal Component Analysis of Electrospray Ionization Fourier Transform Ion Cyclotron Resonance Mass Spectra. *Anal. Chem.* **2013**, *85* (19), 9064−9069.

(7) Tejero, R.; Rossbach, P.; Keller, B.; Anitua, E.; Reviakine, I. Time-of-Flight Secondary Ion Mass Spectrometry with Principal Component Analysis of Titania−Blood Plasma Interfaces. *Langmuir* **2013**, *29*, 902−912.

(8) Ferrari, C.; Foca, G.; Ulrici, A. Handling Large Datasets of Hyperspectral Images: Reducing Data Size without Loss of Useful Information. *Anal. Chim. Acta* **2013**, *802*, 29−39.

(9) Race, A. M.; Steven, R. T.; Palmer, A. D.; Styles, I. B.; Bunch, J. Memory Efficient Principal Component Analysis for the Dimensionality Reduction of Large Mass Spectrometry Imaging Data Sets. *Anal. Chem.* **2013**, *85* (6), 3071−3078.

(10) Genest, S.; Salzer, R.; Steiner, G. Molecular Imaging of Paper Cross Sections by FT-IR Spectroscopy and Principal Component Analysis. *Anal. Bioanal. Chem.* **2013**, *405* (16), 5421−5430.

(11) Jiang, X.; Jiang, Y.; Wu, F.; Wu, F. Quantitative Interpretation of Mineral Hyperspectral Images Based on Principal Component Analysis and Independent Component Analysis Methods. *Appl. Spectrosc.* **2014**, *68* (4), 502−509.

(12) Reese, S. E.; Archer, K. J.; Therneau, T. M.; Atkinson, E. J.; Vachon, C. M.; Andrade, M.; Kocher, J. A.; Eckel-Passow, J. E. A New Statistic for Identifying Batch Effects in High-Throughput Genomic Data That Uses Guided Principal Component Analysis. *Bioinformatics* **2013**, *29* (22), 2877−2883.

(13) Malinowski, E. R. *Factor Analysis in Chemistry*, 3rd ed.; Wiley: New York, 2002.

(14) Anderson, S. L.; Rovnyak, D.; Strein, T. G. Identification of Edible Oils by Principal Component Analysis of ^{1}H NMR Spectra. *J. Chem. Educ.* **2017**, *94* (9), 1377−1382.

(15) Gauthier, J. R.; Burns, D.; Sheng, J.; D'eon, J. C. Exploring the Composition and Authenticity of Honey and Syrup Samples Using Quantitative NMR Spectroscopy and Principal Component Analysis in an Upper-Year Undergraduate Analytical Environmental Course. *J. Chem. Educ.* **2023**, *100* (1), 161−169.

(16) Hupp, A. M. Incorporating Chemometric Methods in the Undergraduate Curriculum: A Problem Set Activity for an Upper-Level Analytical Elective Course. *J. Chem. Educ.* **2023**, *100* (3), 1377−1381.

(17) 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.

(18) De Lorenzi Pezzolo, A. To See the World in a Grain of Sand: Recognizing the Origin of Sand Specimens by Diffuse Reflectance Infrared Fourier Transform Spectroscopy and Multivariate Exploratory Data Analysis. *J. Chem. Educ.* **2011**, *88* (9), 1304−1308.

(19) Sidou, L. F.; Borges, E. M. Teaching Principal Component Analysis Using a Free and Open Source Software Program and Exercises Applying PCA to Real-World Examples. *J. Chem. Educ.* **2020**, *97* (6), 1666−1676.

(20) Friesen, J. B. Forensic Chemistry: The Revelation of Latent Fingerprints. *J. Chem. Educ.* **2015**, *92* (3), 497−504.

(21) Filgueiras, M. F.; Borges, E. M. Quick and Cheap Colorimetric Quantification of Proteins Using 96-Well-Plate Images. *J. Chem. Educ.* **2022**, *99* (4), 1778−1787.

(22) Dumancas, G. G.; Carreto, N.; Generalao, O.; Ke, G.; Bello, G.; Lubguban, A.; Malaluan, R. Chemometrics for Quantitative Determination of Terpenes Using Attenuated Total Reflectance-Fourier Transform Infrared Spectroscopy: A Pedagogical Laboratory Exercise for Undergraduate Instrumental Analysis Students. *J. Chem. Educ.* **2023**, *100* (8), 3050−3060.

(23) Borges, E. M. Hypothesis Tests and Exploratory Analysis Using R Commander and Factoshiny. *J. Chem. Educ.* **2023**, *100* (1), 267−278.

(24) Sequeira, C. A.; Borges, E. M. Enhancing Statistical Education in Chemistry and STEAM Using JAMOVI. Part 2. Comparing Dependent Groups and Principal Component Analysis (PCA). *J. Chem. Educ.* **2024**, *101* (11), 5040−5049.

(25) Feynman, R. P. What is Science, a lecture presented at the fifteenth annual meeting of the National Science Teachers Association, 1966, New York.

(26) Pearson, K. On Lines and Planes of Closest Fit to Systems of Points in Space. *Philos. Mag.* **1901**, *2* (11), 559−572.

(27) Hotelling, H. Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.* **1933**, *24* (7), 498−520.

(28) Seasholtz, M. B.; Kowalski, B. R. The Effect of Mean Centering on Prediction in Multivariate Calibration. *J. Chemom.* **1992**, *6* (2), 103−111.

(29) Shao, L.; Griffiths, P. R. Information Extraction from a Complex Multicomponent System by Target Factor Analysis. *Anal. Chem.* **2010**, *82* (1), 106−114.

(30) Brereton, R. G.; Gurden, S. P.; Groves, J. A. Use of Eigenvalues for Determining the Number of Components in Window Factor Analysis of Spectroscopic and Chromatographic Data. *Chemom. Intell. Lab. Syst.* **1995**, *27* (1), 73−87.

(31) Nadler, B.; Coifman, R. R. The Prediction Error in CLS and PLS: the Importance of Feature Selection Prior to Multivariate Calibration. *J. Chemom.* **2005**, *19* (2), 107−118.
(32) Bro, R.; Smilde, A. K. Principal Component Analysis. *Anal. Methods.* **2014**, *6* (9), 2812−2831.
(33) Abdi, H.; Williams, L. J. Principal Component Analysis. *Wiley Interdiscip. Rev. Comput. Stat.* **2010**, *2*, 433−459.
(34) Kusaka, Y.; Hasegawa, T.; Kaji, H. Noise Reduction in Solid-State NMR Spectra Using Principal Component Analysis. *J. Phys. Chem. A* **2019**, *123*, 10333−10338.
(35) Rutherford, S. H.; Greetham, G. M.; Parker, A. W.; Nordon, A.; Baker, M. J.; Hunt, N. T. Measuring Proteins in H_2O Using 2D-IR Spectroscopy: Pre-Processing Steps and Applications toward a Protein Library. *J. Chem. Phys.* **2022**, *157*, 205102.

https://doi.org/10.1021/acs.jchemed.4c00818
J. Chem. Educ. 2025, 102, 155−163