Published on May 31, 2024 by Arcadia Science **DOI**: 10.57844/arcadia-085e-3ecf

Raman spectroscopy enables rapid and inexpensive exploration of biology

To test its utility in analyzing biological samples, we built an open-source Raman spectrometer and collected spectra from chilis, beer, and algae. We could stratify samples, classify replicates, and link spectra with quantitative traits of beer (ABV) and chilis (perceived heat).

Contributors (A-Z)

Prachee Avasthi, Audrey Bell, Brae M. Bigge, Ben Braverman, Tara Essock-Burns, Megan L. Hochstrasser, Ryan Lane, Cameron Dale MacQuarrie, David G. Mets, Austin H. Patton, Dennis A. Sun, Harper Wood, Ryan York

Version 3 · Mar 31, 2025

Purpose

Raman spectroscopy is a non-destructive technique that provides a unique chemical fingerprint based only on the interaction of light with a sample. It’s been used extensively in materials science applications and more recently, in biology. This technique doesn’t require molecular or chemical labeling (it’s “label-free”), making it a potentially useful tool for studying organisms without genetic tools.

We wondered if we could build a Raman spectrometer using open-source protocols and use it to rapidly distinguish samples based on chemical properties in a label-free way, with minimal data processing. We decided to try a hackathon to test this idea — we selected three types of samples (beer, chilis, and algae) and found that the spectra were reproducible and had sufficient dynamic range to do comparative analyses. We were able to use the Raman spectra to differentiate the three types of samples and to distinguish subgroups of samples within a given type. Beer sample spectra varied by alcohol content and by type. Chili pepper data clustered by perceived heat (Scoville units) and color. We could differentiate algae by genetic background. Finally, we found that specific spectral regions correlate with quantitative characteristics of beer (alcohol by volume) and chilis (perceived heat).

Our work highlights the utility and ease of this technique. We hope it will empower scientists to capture the chemical composition of samples and extract a great degree of high-dimensional data from Raman spectra. We imagine this report could also be useful for science educators who want to use the OpenRAMAN resource and our code to run a lab class on Raman spectroscopy. We'd love to know if you try this technique and whether it allows you to distinguish features in a way that isn't possible or is more difficult using other methods.

- All associated **code** for analyzing the spectral data is available in this GitHub repository.
- **Data** from this pub, including the raw spectra of beer samples, chili peppers (seeds and flesh), and algal samples, are available in the "data" folder of the GitHub repo.
- The **comprehensive parts list** that we used to build the OpenRaman is in the "resources" folder of the GitHub repo.

Background and goals

At Arcadia, we're mapping genetic and phenotypic diversity across the tree of life to aid in predictive modeling and biological discovery. We've recently shown that high-dimensional phenotyping can improve the accuracy of phenotypic models **[1]** and, likely, genotype-to-phenotype mappings. However, measuring high-dimensional phenotypes is often laborious, most studies only measure one phenotype, and phenotyping often requires you to know what you're looking for by pre-selecting a specific phenotype to quantify. In this pub, we evaluate the suitability of Raman spectroscopy for high-throughput, high-dimensional agnostic phenotype acquisition.

Raman spectra capture information about the chemical composition of a sample. Samples are briefly exposed to a high-intensity, single-wavelength light source. Most of the light is reflected or scattered elastically and is the same wavelength as the incident light. A minor fraction of the scattered light shifts wavelength. These shifts are caused by energy loss through vibrational or rotational absorption and shifts are characteristic of specific chemical bonds. Thus, the spectral distribution and intensity of this inelastically scattered light provide a fingerprint for the chemical bonds in the sample **[2]**.

Raman spectroscopy of cells has recently been shown to contain holistic proteomic **[3]** and expression **[4]** data. In these studies, the authors used cellular Raman spectra to predict entire proteomes and single-cell expression profiles. Furthermore, we've shown that spectra of differing species reflect their phylogenetic relationships **[5]**.

To better evaluate the utility of Raman spectroscopy for the analysis of biological information, we conducted a two-day hackathon **[6]** where we used a Raman spectrometer (OpenRAMAN) that we built in preparation to collect spectra for three types of biological samples (beer, chili peppers, and algal species). We then looked to see if we could 1) use the spectra for clustering/classification and trait/feature prediction, and 2) identify the importance of specific wavelengths for these predictive tasks. We selected samples that were likely to have clear and quantifiable dimensions of variation, such as alcohol content for beer and perceived heat for chili pepper.

Raman spectra contain enough information to not only differentiate samples but also to differentiate sample types based on combinations of features. Skip straight to these results or continue reading to review our methodology.

> **SHOW ME THE DATA**: Data from this pub, including the raw spectral data, are available here (DOI: 10.5281/zenodo.11406248).

The approach

We ran an internal hackathon to quickly assess the utility of Raman spectroscopy in analyzing complex biological samples. Hoping to answer this question in just a few days, we chose a low-cost, open-source spectrometer to build ahead of time and test during the hackathon (OpenRAMAN). We designed our experiment to test three types of samples with varying attributes that we expected could be differentiated by their Raman spectra. We selected beer with varying levels of alcohol content (ethanol) and of different varieties representing different brewing yeasts, hops, malt, and other ingredients. We chose chilis that ranged in capsaicin level, color, and state (fresh vs. dried). Finally, we used algae species of varied genetic backgrounds that we were already using in other projects **[7]**.

Building an open-source Raman spectrometer

We built our Raman spectrometer using instructions from OpenRAMAN and YouTube (Figure 1).

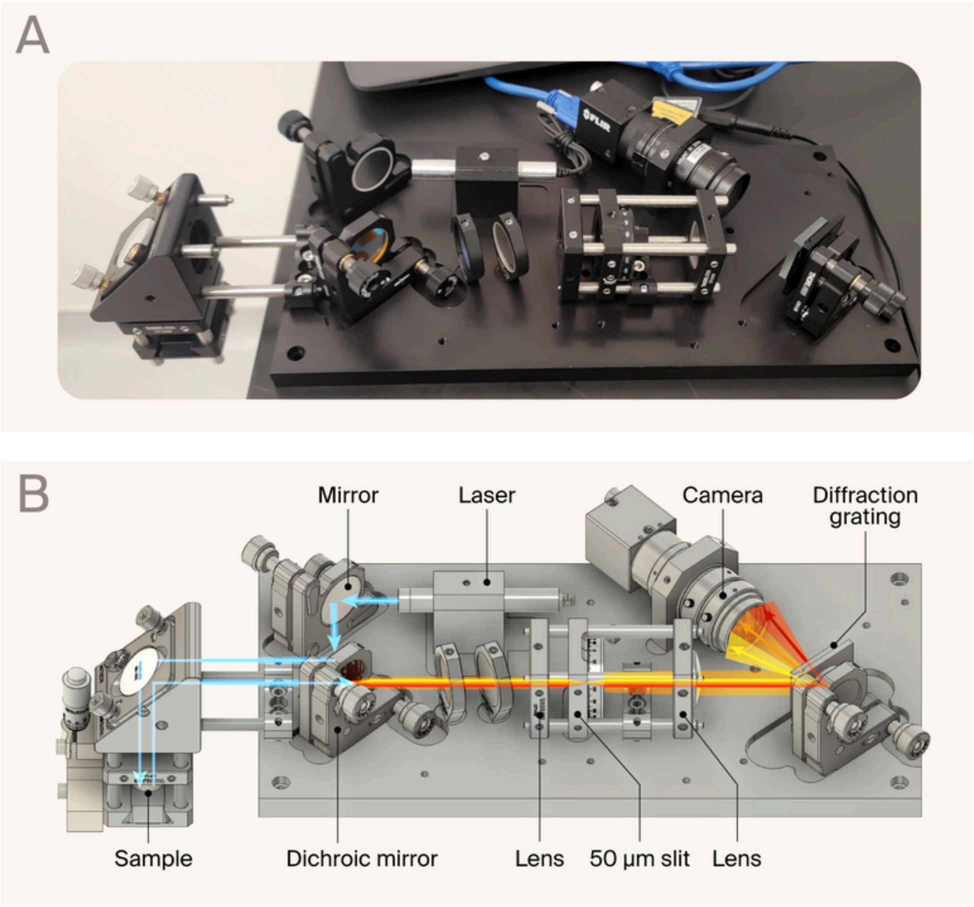


Figure 1

Configuration of the OpenRAMAN spectrometer.

(A) Photograph of the assembled OpenRAMAN spectrometer.

(B) Corresponding schematic with labeled parts and the path of green, yellow, orange, and red light. We used a green laser (532 nm), but we've depicted its path in blue to make the image color-blind-friendly.

We built our spectrometer according to the directions for the "Starter Edition" with a few minor changes. Namely, we made the 3D-printed components using inexpensive fused deposition modeling instead of the suggested selective laser sintering due to tool availability. We also modified the inner diameter of the camera bracket from 32 mm to 34 mm to accommodate our camera lens. Finally, the 550 nm dichroic mirror was not available, so we replaced it with a 567 nm dichroic mirror (Thorlabs DMLP567). For ease of communication with our analysis computer, our camera (Teledyne Flir BFS-

U3-16S2M-CS) used a universal serial bus 3 interface instead of a gigabit ethernet interface.

We’ve put together a comprehensive parts list that includes all the parts we used, plus other necessary tools and materials, which you can find here:

CSV

`OpenRaman starter edition (comprehensive BOM) - 2020-06_3 - STARTER EDITION ASSY.csv`

Download

Data collection and sample preparation

From the options available at Berkeley Bowl West (Berkeley, CA, USA), we selected a variety of beers differing in alcohol content (alcohol by volume, ABV) and style. We collected the characteristics of these beers from both brewery webpages and the beer information aggregation website Untappd. These data reflect the values as of March 21st, 2024; given their crowdsourced origin, they’re likely to change over time. For sample preparation, we poured beer into weigh boats, where we agitated the beer to reduce bubbles and carbonation before pipetting 5 μl of each sample onto Parafilm and placing it in the sample chamber of the spectrometer.

Icon	Beer	Brewery	ABV (%)	Style	IBU	Untappd rating (out of five)	Unt tag
🍺	Dark Majik	Lough Gill (Sligo, Ireland)	11.0	Imperial Irish oatmeal coffee stout	0	3.91	Cof Smo Rich Boo
🍺	Sneaky AF	Del Cielo Brewing Co (Martinez, CA)	10.0	Triple IPA	0	3.99	Bala Aron Ligh Bod Eart
🍺	Big Love	Almanac (Alameda, CA)	9.0	Hazy double IPA	0	3.88	Hop Citr Smo Stro Bod
🍺	Gnomes Gone Rogue	Original Pattern (Oakland, CA)	8.1	Hazy double IPA	0	4.13	Haz Hop Pine Mal
🍺	Otto's Jacket	Cellarmaker (Oakland, CA)	7.0	West Coast IPA	58	3.99	Mal Smo Car Swe
🍺	Kimchi Sour	Dokkaebier (Oakland, CA)	6.6	Sour	14	3.56	Ligh Bod Cris Citr Hop Grai
🍺	Love	Almanac (Alameda, CA)	6.1	Hazy IPA	0	3.88	Ging Ligh Bod Spic Smo
🍺	Colour Me Murphy	Original Pattern Brewing (Oakland, CA)	6.0	Irish red ale	0	3.86	Clea Citr Ora Hop

Icon	Beer	Brewery	ABV (%)	Style	IBU	Untappd rating (out of five)	Unt tag
🍺	Hunky Jesus	Laughing Monk (San Francisco, CA)	5.5	Blood orange pale ale	0	3.71	Dar Sm Cof Flat
🍺	Kolschtastic	Gigantic Brewing Co (Portland, OR)	5.2	Kolsch	25	3.60	Ligh Bod Cle Swe Hop
🍺	Temescal Pils	Temescal Brewing (Oakland, CA)	5.0	Pilsner	3.71	3.71	Ligh Bod Cle Swe Hop
🍺	Helles (Long Nights Edition)	Wayfinder (Portland, OR)	4.9	Lager	20	3.90	Ligh Bod Cle Brig
🍺	Even MORE IRISH Jesus	Evil Twin Brewing (North Haven, CT)	4.7	Dry stout	0	3.64	Dar Sm Cof Flat
🍺	Party Wave	Headlands (Lafayette, CA)	4.2	Light lager	14	3.83	Ligh Bod Sm Effe Stra Wat

Table 1

Beer varieties sampled.

We selected 20 chili peppers from Berkeley Bowl West (Berkeley, CA, USA), aiming for a wide distribution of spiciness and color. We dissected fresh and dried whole chili pepper varieties into two different sample types (flesh and seed) using razor blades on aluminum foil. Crushed red pepper flakes contain both seeds and flesh, so we selected a fragment of flesh and a fragment of seed for testing. We cut the flesh into roughly 0.5 cm^3 pieces and collected spectra from the interior face. We found that spectra from whole seeds were qualitatively similar to dissected seeds, so we're presenting only spectra captured from whole seeds here, but included the data acquired from the pepper flesh in our GitHub repo. We used forceps to transfer pepper samples onto Parafilm for data collection.

Icon	Chili variety (as labeled at Berkeley Bowl)	Abbreviation in GitHub repo (arbitrarily assigned)	Chili condition	Perceived heat range (Scoville units)	Median Scoville units
🫑	Green bell	GrBe	Fresh	0	0
🌶️	Red Thai	ReTh	Fresh	110,000	110,000
🫑	Hot Italian frying	HoIt	Fresh	100-1,000	550
🫑	Poblano	Pbl	Fresh	1,000-1,500	1,250
🫙	Ancho (dried poblano)	Ancho	Dried	1,000-1,500	1,250
🌶️	Hungarian wax	HuWa	Fresh	1,000-15,000	8,000
🫑	Chilaca	Chil	Fresh	1,000-2,500	1,750
🌶️	Serrano	Serr	Fresh	10,000-23,000	16,500

Icon	Chili variety (as labeled at Berkeley Bowl)	Abbreviation in GitHub repo (arbitrarily assigned)	Chili condition	Perceived heat range (Scoville units)	Median Scoville units
🫙	Chili de arbol	Arbol	Dried	15,000-30,000	22,500
🫑	Orange habañero	OrHa	Fresh	150,000-350,000	250,000
🌶️	Red Fresno	Fres	Fresh	2,500-10,000	6,250
🫑	Jalapeño	Jala	Fresh	2,500-8,000	5,250
🫙	Chipotle (dried jalapeno)	Chip	Dried	2,500-8,000	5,250
🫑	Indian long	InLo	Fresh	25,000-100,000	62,500
🫙	Crushed red	CrRe	Dried	32,000-48,000	40,000
🫑	Shishito	Shis	Fresh	50-200	125

Icon	Chili variety (as labeled at Berkeley Bowl)	Abbreviation in GitHub repo (arbitrarily assigned)	Chili condition	Perceived heat range (Scoville units)	Median Scoville units
🌶	Anaheim	Anah	Fresh	500–2,500	1,500
🌶	Yellow wax	YeWa	Fresh	5,000–15,000	10,000
🌶	Green Thai	GrTh	Fresh	50,000–100,000	75,000
🫙	New Mexico	NeMe	Dried	800–1,400	1,100

Table 2

Pepper varieties and phenotypes.

We collected spectra for both flesh and seed for each sample, but only present data for the seeds here. All chili pepper samples are cultivars within the species *Capsicum annuum* except the orange habañero (*Capsicum chinense*).

We collected spectra from several unicellular algae, including freshwater *Chlamydomonas reinhardtii*, *Chlamydomonas smithii*, four hybrid strains from crossing these species **[7]**, and the marine alga *Isochrysis galbana* (Table 3). Using sterile loops, we transferred algae from solid media culture plates to Parafilm for data collection.

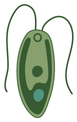
Icon	Species	Strain	Source	Medium (with 1.5% agar)
	Chlamydomonas reinhardtii	cc-124	CRC	Tris-acetate-phosphate (TAP)
	*Chlamydomonas smithii*	cc-1373	CRC	TAP
	Chlamydomonas hybrids	ACDC 13F3, 13F4, 13F5, 13F6	Arcadia Science from the Arcadia Chlamydomonas Diversity Collection (ACDC)	TAP + yeast extract (0.4%) + carbenicillin (500 mg/L)
	Isochrysis galbana	UTEX LB 987	UTEX	Erdschreiber’s

Table 3

Algal types sampled.

Data analysis

We clustered spectra using linear dimensionality-reduction methods. First, we performed unsupervised clustering of the full spectral dataset via principal component analysis (PCA). We assessed sample relationships by comparing the first two principal components (Figure 3). We then used linear discriminant analysis (LDA) to assess the extent to which we could classify individual samples within each data class (beer, chilis, algae). For each, we used the `lda` function in the R package MASS **[8]** to find a linear combination of spectral features that best classified samples (i.e., beer type, chili variety, and algal species). We assessed each LDA by comparing the first two linear discriminants (Figure 4).

Next, we assessed the extent to which we could identify regions of these spectra that correlate with quantitative features of different beers or chilis. Specifically, we examined the alcohol content of each beer (ABV), and, independently, the perceived heat of each chili (Scoville units). We obtained ABV values from each beer can (Table 1) and Scoville units from several sites, including Wikipedia, Bonnie Plants, Chili Pepper Madness, and Scoville Scale (Table 2). In cases where a chili variety had a range of reported Scoville values, we used the median. The distribution of Scoville units was highly skewed, so we transformed the data so that we could perform analyses that assume a normal distribution. We added one to all Scoville values to eliminate zeros and transformed these measures using $\log_{10}$. For each sample, we collected between two and four spectra. We used the median of these spectra for subsequent analyses.

We expect that many of the components of these spectra will not be useful in predicting any particular quantitative feature of the samples. We, therefore, chose the least absolute shrinkage and selection operator (LASSO) regression **[9]** as implemented using the glmnet R package (version 4.1.8) **[10]**. Unlike the ordinary least squares solution to regression problems, this method is regularized using the L1 norm and expects that few model parameters contribute to a trait.

LASSO has a single tunable parameter, the L1 penalty (or λ), that determines the degree of regularization. To identify a value of λ that leads to the most usefully predictive model, we took a permutation-based approach. For 5,000 permutations, we randomly subsampled 75% of our data. We then used this 75% to tune λ through cross-validation (according to **[10]**). We tested the predictions for each permutation on the 25% of data that we didn’t use in the training. Following all permutations, we then used the λ that resulted in the most accurate predictive model to train a final model using all of the data. For significance testing, we calculated confidence intervals for each spectral position (pixel) from these permutations. We considered each location significant at $p < 0.05$. We note that these are local statistical tests that do not account for the multiple tests conducted in this study. The coefficients resulting from that final model are those presented in Figure 5 and Figure 6.

All **code** generated and used for the pub is available in this GitHub repository (DOI: 10.5281/zenodo.11406248), including scripts and notebooks used for processing and visualizing the data.

Additional methods

We used ChatGPT to help write code and add comments to our code. We also used it to generate the average length and typical uses of the peppers in Table 2.

The results

SHOW ME THE DATA: Access our raw Raman spectral data here.

Raw spectra are reproducible across technical replicates

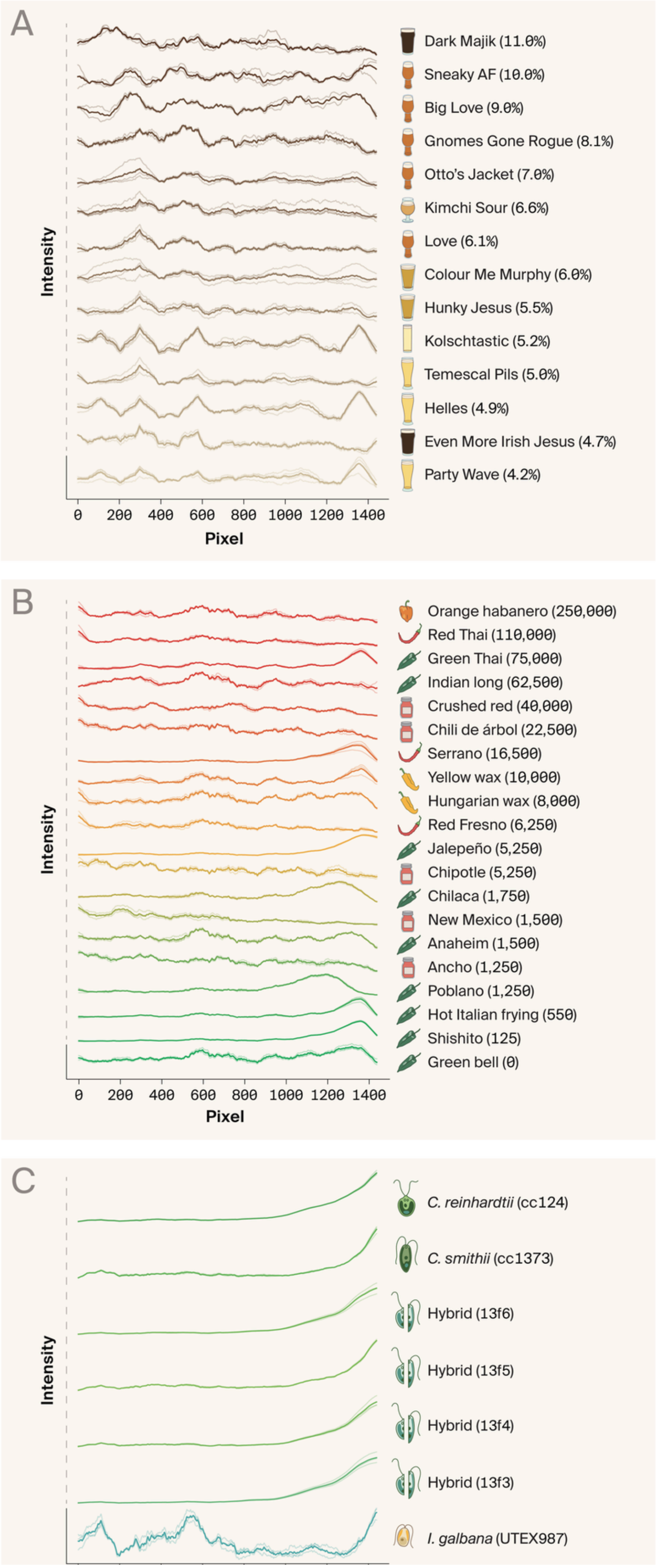
Since spectroscopic measurements can be influenced by various noise sources — sample heterogeneity, hardware variability, fluorescence — we were interested in qualitatively assessing how consistently our spectra performed before more complex analyses (Figure 2). Encouragingly, spectra were similar within sample type (e.g., within beer or chilis) and reproducible across technical replicates (Figure 2). Furthermore, the spectra differed across sample types (Figure 2). Some of these differences seemed to reflect readily apparent features of the samples. For example, samples with “greener” color (green/yellow chilis and algae) seemed to have increased spectral intensity in the 1200–1400 pixel (px) region (consistent with chlorophyll fluorescence; Figure 2, B–C). Similarly, light beers displayed a spectral peak between 1,300–1,400 px that other beer types lacked (Figure 2, A). We concluded that our measurements were sufficiently consistent, and displayed enough dynamic range across samples, that quantitative analyses would be interesting to pursue.

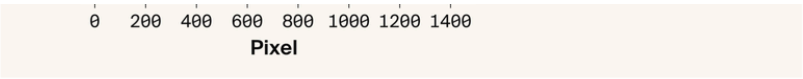
Clustering the spectra lets us separate samples by type

A potential benefit of Raman spectroscopy is that a single rapidly acquired measurement may provide enough information to classify complex biological samples. We explored this possibility by performing unsupervised clustering via principal component analysis (PCA) on raw spectra. We reasoned that the outcome of the PCA could inform us about the structure and richness of information contained within the spectra. For example, if we observed extreme mixing of samples among the principal components (i.e., no clustering), then we might conclude that the spectra are either too complex or too noisy to easily identify samples from raw measurements. On the other hand, if we found tight clusters corresponding to sample type, then spectra may be highly sample-specific but lack enough quantitative information to usefully stratify similar samples based on their biochemical differences.

Comparing the first two principal components, we qualitatively found that samples largely clustered by type and that we could separate them linearly (Figure 3). For example, PC1 appeared to mostly separate algae from the other samples, while PC2 delineated beer from chilis (Figure 3). Sample types also displayed qualitatively differing amounts of variation. Algae samples were the most variable, followed by beer and then chilis (Figure 3). These findings suggest that our spectra fall in between the two extremes outlined above: they contain enough information to cluster sample types, but there's also measurable variation within the different sample types (i.e., beer, chilis, and algae). This encouraged us to explore the nuances of spectral data within sample types.

We were interested to see how a classifier might perform when applied to our





spectra. Specifically, we created linear classifiers predicting each sample type from spectra via linear discriminant analysis (LDA). We found that, in each case, the first two linear discriminants grouped technical replicates together. Individual beer samples did cluster approximately according to their alcohol content — the three highest-ABV beers clustered together, including Dark Majik at 11%, Sneaky AF at 10%, and Big Love at 9% (Figure 4, A). Interestingly, though two of these three are IPAs, similar-style beers like a second Hazy Double IPA did not join this cluster. We also found that three of the lighter-style beers with lower alcohol content clustered together, including the Kolsch Kolchstastic at 5.2%, the lager Helles (Long Nights Edition) at 4.9%, and the light lager Party Wave at 4.2% (Figure 4, A). The key exception was the pilsner, Temescal Pils (5.0%), which did not cluster with the other lighter-style, low-alcohol beers. Instead, the pilsner joined the third cluster, which includes beers with an intermediate ABV (Figure 4, A). The chili seed samples tended to be sorted by color of the chili on LD1, with the red chilis and the various dried chilis to the left and the green chilis to the right (Figure 4, B). Across samples, one of the dominant signals was pigment fluorescence, including chlorophyll and carotenoids. This held true even for chili seeds. Finally, we found that each algal sample clustered independently,

Figure 2

Raman spectra of beer samples, chili pepper seeds, and algae.

(A) We've ordered beers and their spectra by alcohol content, with the highest ABV at the top.

(B) We've ordered chili pepper seeds and their spectra by perceived heat/spiciness (Scoville units), with the hottest at the top.

(C) For algae spectra, we've listed the two parent species (*Chlamydomonas reinhardtii* and *C. smithii*) first, then the hybrids from the genetic cross, and then a more distantly related alga, *Isochrysis galbana*.

The mean spectrum for each sample (bold line) is the average of two to four measurements (lighter lines) and is shown as intensity (y-axis) across pixels (x-axis). The y-axis for each spectrum is automatically scaled in each plot to show the full range of intensity values.

demonstrating that the cross between *Chlamydomonas reinhardtii* and *Chlamydomonas smithii* resulted in unique progeny that are differentiable from either parent (Figure 4, C). This suggests that the genetic and resultant physiological and chemical differences between these unique hybrid strains are captured in Raman spectra. These spectra can be used as high-dimensional phenotypes to differentiate both species and strains and potentially improve genotype-to-phenotype mappings **[1]**.

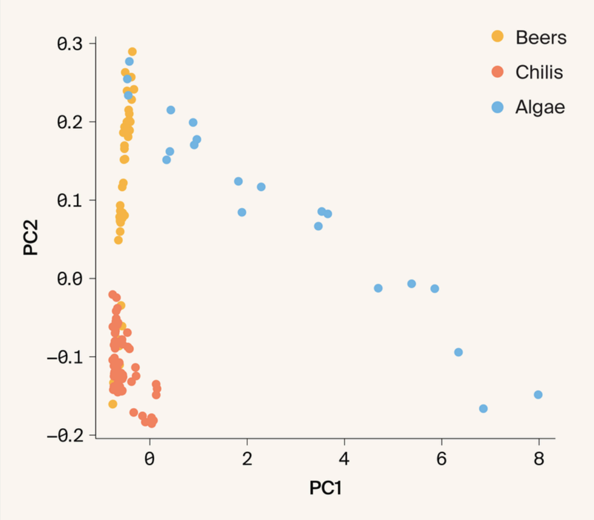


Figure 3

Spectral clustering of samples via principal component analysis (PCA).

Clustering of the full dataset using the first two principal components.

Specific regions of the spectra correlate with quantitative features of the samples

Our clustering results show that these Raman spectra contain sufficient information to identify individual biological samples, suggesting they might also contain information about quantitative features that varied across those same samples. To test this possibility, we identified spectral regions that significantly capture information about beer alcohol content (ABV) and the perceived heat of a chili (Scoville units). We didn’t analyze quantitative traits for algae because we tested fewer individual samples (i.e., strains) than we did for chilis and beer. For both ABV (Figure 5) and Scoville units (Figure 6), we conducted a LASSO, a regularized form of regression, where intensities at individual spectral positions were independent variables and the quantitative trait was the dependent variable. We chose LASSO because it’s effective in cases where only very few of the model parameters (intensity at individual pixels in the spectra) influence the response variable, something we expect to be true for these data. We optimized our model for the prediction of “test” data not used during training. Therefore, significant spectral features are predictive of the particular quantitative trait.

We determined the significance of each spectral position by permutation test (see “Data analysis” for details).

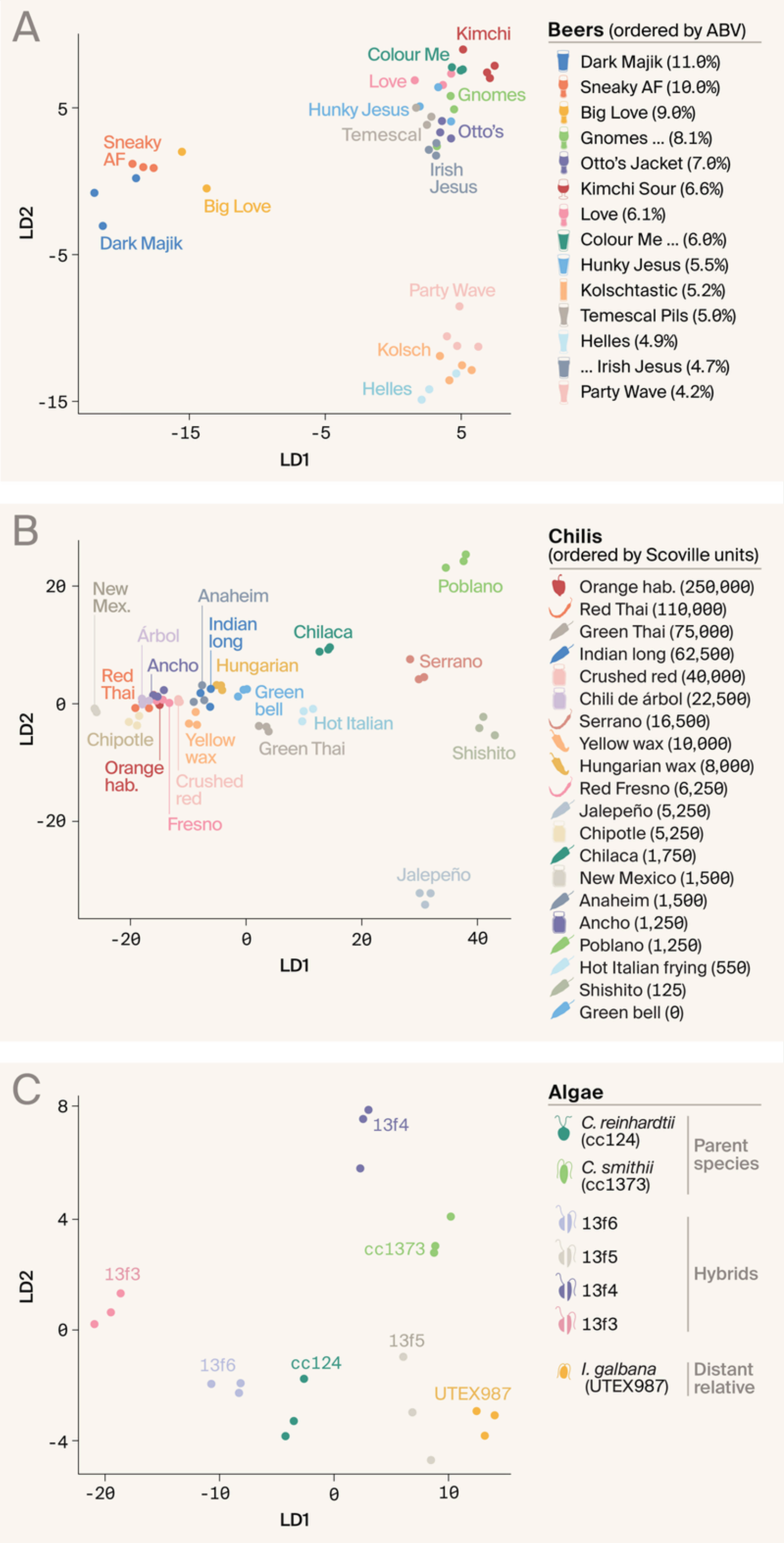


Figure 4

Spectral clustering of samples via linear discriminant analysis (LDA).

2D clustering within LDA space for (A) beers, (B) chilis, and (C) algae.

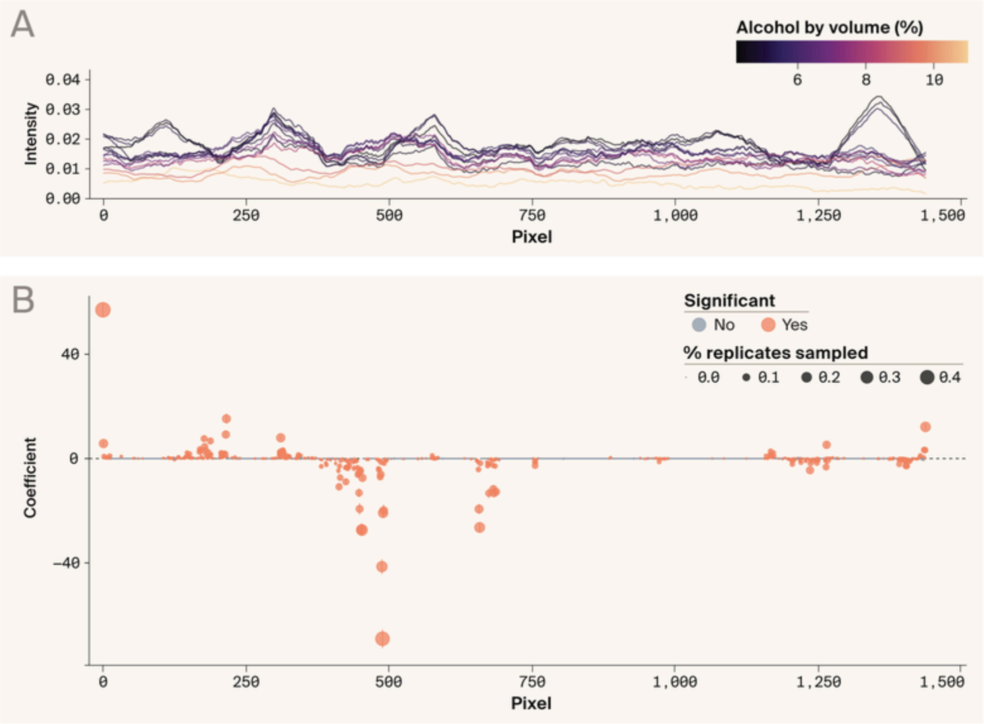


Figure 5

Local importance and contribution of Raman spectra in predicting alcohol content (ABV) of beers, as inferred using LASSO regression.

(A) Each line shows the mean spectrum for a specific beer. The color of the line corresponds to the ABV for each beer.

(B) Each point corresponds to a single pixel along the spectrum, and its position along the y-axis corresponds to how strongly spectral intensity at that position predicts ABV. Positive coefficients indicate spectral positions that positively predict ABV, whereas those with negative coefficients negatively predict ABV. The size of points corresponds to the percentage of bootstrap replicates (n = 5,000) in which that spectral position was retained by L1 regularization (LASSO) regression; vertical lines associated with each circle indicate the 95% confidence intervals for each inferred coefficient. Points in orange are those for which the bootstrapped 95% confidence intervals are non-overlapping with zero.

Our analysis of beer samples identified several regions of Raman spectra that significantly predict ABV (Figure 5, bootstrapped confidence intervals, $p < 0.05$). Although the LASSO regression treats each spectral position as independent of the others, the spectral positions with significant coefficients appear (qualitatively) to cluster in spectral space, though we did not formally test this. For instance, the major peaks in spectral intensity for lower-ABV beers are often flanked by spectral positions with significant coefficients (Figure 5, B). There are apparent clusters of significant coefficients at these positions, where the intensity of Raman signal begins to shift. Thus, we can use these spectra to identify features that significantly predict the ABV of a sample.

Across the chili seed samples, chlorophyll fluorescence drove much of the variation (Figure 6, A, pixels 1,200–1,440). Despite this, we identified spectral regions that predict perceived heat (Figure 6, B; bootstrapped confidence intervals, $p < 0.05$). The regression coefficients for spectral regions with variation driven by chlorophyll or carotenoid fluorescence (Figure 6, B; pixels 1,200–1,440) are much smaller than coefficients for other sections of the spectra. This pattern could indicate that chemicals causing Raman shifts in this spectral range contribute less to a pepper's perceived heat than chemicals causing Raman shifts in other spectral ranges. Alternatively, it could be that the strong chlorophyll or carotenoid fluorescence reduces our ability to estimate the contribution of truly meaningful features. A less exploratory study would benefit from more rigorous control of this confounder. One could explore this further by comparing the spectral data from seeds to flesh and isolating the spectral contribution of the pigment (chlorophyll and carotenoids). Though not presented here, our data from the analysis of chili flesh samples are also available in our GitHub repository.

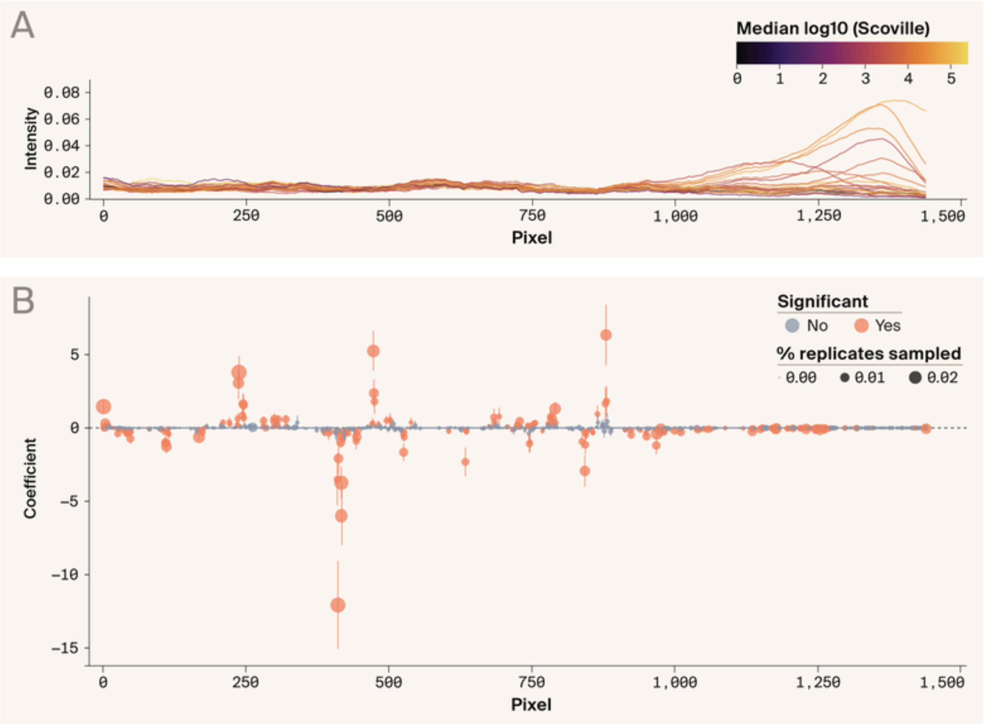


Figure 6

Local importance and contribution of Raman spectra in predicting perceived heat of peppers ($\log_{10}$-transformed Scoville units), as inferred using LASSO regression.

(A) Each line shows the mean spectrum for a specific chili seed sample. The color of the line corresponds to the log-transformed Scoville units for each chili pepper.

(B) Each point corresponds to a single pixel along the spectra, and its position along the y-axis corresponds to how strongly spectral intensity at that position predicts perceived heat. Positive coefficients indicate spectral positions that positively predict perceived heat, whereas those with negative coefficients negatively predict perceived heat. Size of points corresponds to the percentage of bootstrap replicates (n = 5,000) in which that spectral position was retained by L2 regularization (LASSO regression); vertical lines associated with each circle indicate the 95% confidence intervals for each inferred coefficient. Points in orange are those for which the bootstrapped 95% confidence intervals are non-overlapping with zero.

The analyses of both beer and chilis show that these spectra contain information about quantitative features of these biological samples and we can identify the components of the spectra that contribute to these features.

Key takeaways

1. Raman spectroscopy yields meaningful data about the chemical composition of biological samples, and there's a cheap, quick, easy, and open-source way to build your own Raman spectrometer (OpenRAMAN).
2. Testing the OpenRAMAN spectrometer on chilis, beer, and algae showed that this approach is sufficient to classify samples by their spectra and associate them with quantitative traits.
3. High-dimensional phenotyping through Raman spectroscopy is useful and accessible.

Next steps

In this pub, we rapidly tested the feasibility of using a tool for our downstream work by running a hackathon. This hackathon structure was quite useful for constraining a small project in time and scope and we'll likely try it again in the future. Because of the ease of data collection and application of machine learning algorithms, we'll continue to leverage Raman spectroscopy, including using the inexpensive OpenRAMAN spectrometer, as a powerful approach for probing biology. We'd like to help make Raman spectra from biological samples easier to interpret, so we'd love to hear if there are any Raman-focused FAIR databases that would be appropriate for these spectra. We've shared our data in the GitHub repo associated with this pub, but it would be great to make them more discoverable and contribute to a shared, centralized resource.

References

1 Avasthi P, Mets DG, York R. (2024). Harnessing genotype-phenotype nonlinearity to accelerate biological prediction. https://doi.org/10.57844/ARCADIA-5953-995F

2 Jones RR, Hooper DC, Zhang L, Wolverson D, Valev VK. (2019). Raman Techniques: Fundamentals and Frontiers. https://doi.org/10.1186/s11671-019-3039-2

3 Kamei KF, Kobayashi-Kirschvink KJ, Nozoe T, Nakaoka H, Umetani M, Wakamoto Y. (2023). Revealing global stoichiometry conservation architecture in cells from Raman spectral patterns. https://doi.org/10.1101/2023.05.09.539921

4 Germond A, Ichimura T, Horinouchi T, Fujita H, Furusawa C, Watanabe TM. (2018). Raman spectral signature reflects transcriptomic features of antibiotic resistance in Escherichia coli. https://doi.org/10.1038/s42003-018-0093-8

5 Avasthi P, Patton AH, York R. (2024). Raman spectra reflect complex phylogenetic relationships. https://doi.org/10.57844/ARCADIA-X8WK-SF94

6 Medina Angarita MA, Nolte A. (2020). What Do We Know About Hackathon Outcomes and How to Support Them? – A Systematic Literature Review. https://doi.org/10.1007/978-3-030-58157-2_4

7 Avasthi P, Braverman B, Essock-Burns T, Garcia G, MacQuarrie CD, Matus DQ, Mets DG, York R. (2024). Phenotypic differences between interfertile Chlamydomonas species. https://doi.org/10.57844/ARCADIA-35F0-3E16

8 Venables WN, Ripley BD. (2003). Modern applied statistics with S, 4th ed. https://cran.r-project.org/web/packages/MASS/index.html/

9 Tibshirani R. (1996). Regression Shrinkage and Selection Via the Lasso. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

10 Friedman J, Hastie T, Tibshirani R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. https://doi.org/10.18637/jss.v033.i01