JOURNAL OF CHEMICAL EDUCATION

pubs.acs.org/jchemeduc

Laboratory Experiment

Integrating Data Science and Machine Learning with an Aldol Condensation Laboratory

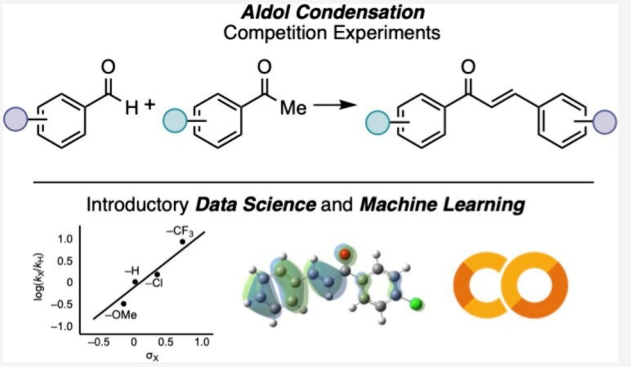
Daniel S. Min,[‡] Flora Fan,[‡] and Abigail G. Doyle*

Cite This: https://doi.org/10.1021/acs.jchemed.5c00994

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: We report the development of an undergraduate organic chemistry laboratory to introduce students to modern applications of data science tools and machine learning algorithms in organic chemistry. Data science and machine learning have become increasingly applied to organic chemistry systems built upon physical organic principles of reactivity to better analyze and interpret data. Given that postexperimental analysis is central to any scientific study, we envision that the incorporation of these techniques at an introductory level into the undergraduate chemistry education curriculum will be invaluable in exposing students to contemporary research tools and working with shared data. Herein we describe a two-part experiment, using the experimentally straightforward Claisen−Schmidt aldol condensation reaction with commercially available reagents, to introduce concepts of computational featurization and data processing for multivariate linear regression models at the undergraduate level that can easily be incorporated into organic instructional laboratories.



KEYWORDS: *Upper-Division Undergraduate, Interdisciplinary, Laboratory Instruction, Organic Chemistry, Collaborative Learning, Catalysis, Computational Chemistry, Machine Learning, Data Science*

INTRODUCTION

Artificial intelligence (AI) tools have rapidly become integral across a wide range of disciplines from finance and healthcare to science and education. Specifically, machine learning (ML), a subset of AI that learns through data, is becoming increasingly prevalent, as seen through examples of ChatGPT, photo editing, and language translation. In chemistry, ML has seen a wide array of applications, including reaction prediction,[1−4] condition optimization,[5,6] retrosynthesis pathway prediction,[7,8] and mechanistic investigation.[9−11] Across these areas, ML has enhanced chemists' ability to perform experiments faster and understand resulting reactivity that may be outside intuition.[12] In particular, ML algorithms have been combined with established physical organic reactivity principles, using extracted electronic and steric features of reagents, catalysts, and/or ligands to build predictive and mechanistically interpretable linear free energy reactivity and selectivity models.[13−15] Due to this growing importance and prevalence of ML in chemistry, it is worthwhile that students are introduced to the concepts and applications of machine learning in their undergraduate studies.[16−18] Currently, published lab modules that incorporate machine learning or computational tools rarely concurrently expose the students to the wet-lab experimental counterpart,[19−22] leading to a disconnect in understanding the machine learning pipeline. For example, using pre-existing, curated datasets for analysis may obscure key processes in data collection and interpretation of chemical data, such as how experimental error affects modeling or human biases introduced during synthesis.[23,24]

We sought to develop a lab module that is more cohesively integrated with subsequent computational analysis, with the goal of introducing students to modern directions and workflows in chemistry research. In this laboratory module, students practice experiential learning[25] by completing an end-to-end project starting from wet-lab experiments to data analysis and ML predictions. Unlike a typical undergraduate organic chemistry laboratory, students gathered individual pieces of data through different experiments and then directly collaborated with their peers to use the collective pooled data to perform their analysis (see Figure 1), the process of which encouraged higher-order thinking.[26,27]

OBJECTIVES

This two-part laboratory experiment aims to introduce upper-division undergraduate chemistry students to physical organic

Received: July 17, 2025
Revised: January 24, 2026
Accepted: January 29, 2026

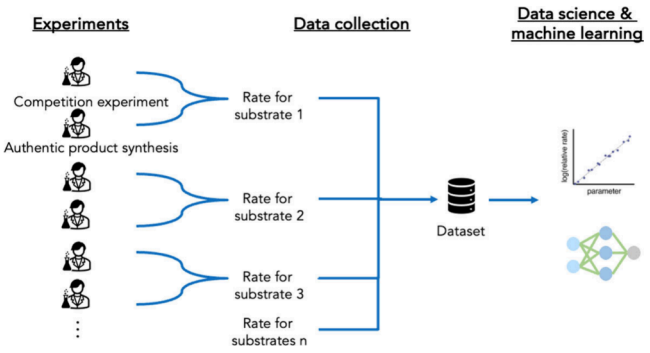


Figure 1. Workflow for this laboratory.

chemistry principles as well as modern data science and machine learning tools to analyze reactivity patterns. The laboratory is designed to be accessible to students with little to no prior programming experience. At the end of this instructional laboratory experiment, students were expected to

1. work together to gather experimental data using standard organic chemistry laboratory techniques, including reaction setup and workup, as well as learning how to use gas chromatography (GC) for analysis;
2. construct linear free energy relationships, in particular Hammett plots,[28,29] to model the electronic effects of substrate substitution in a condensation reaction;[30,31]
3. learn how to use a curated Google Colab notebook and process a spreadsheet of experimental and computational data;
4. demonstrate the potential of data science and machine learning tools to afford predictive and interpretable models on more varied datasets than are typically applicable to a Hammett study of electronic substituent effects.

SOFTWARE DETAILS

The goal of the module is to introduce undergraduate students to how various computational methods, such as basic programming, data science, machine learning, and quantum-mechanical calculations, can be applied to organic chemistry.

The code is hosted on a web-based free cloud service, Google Colab, which requires minimal setup requirements and is agnostic to computer types. Furthermore, as Google Colab is compatible with notebook files (.ipynb), it allows for integration of written notes and images to aid the students' learning. Unlike a traditional python script file (.py), it allows for sectionwise execution of the code, allowing the instructor to break down the large codebase.

Python was chosen as the primary language for three main reasons: (1) the readability and flexibility of python makes the language beginner-friendly; (2) the dominance of python in machine learning in industry and academia allows for a directly applicable language for students; and (3) the abundance of available libraries allows for a modular codebase, abstracting often long and unintuitive functions. The libraries used are listed in Table 1.

OVERVIEW OF THE LABORATORY EXPERIMENT

This experiment is designed for upper-division undergraduate chemistry majors (juniors and seniors) but may also be suitable for sophomore organic chemistry laboratories. Previous programming or computer science knowledge is not required, although it may be helpful. The module requires 4−5 hours to complete, which may take place over two 2−3 hour laboratory periods: in the first, the students perform the reaction in the laboratory to gather experimental data, and in the second, they analyze their data together as a class and construct reactivity models. Students completed a prelaboratory worksheet to demonstrate comprehension of the experiment they would be running and the associated safety precautions. After the completion of the experiment and data analysis, they completed a postlaboratory worksheet to review their understanding of the content.

Table 1. Python libraries imported in Google Colab Notebook

Library Name	Description
Matplotlib, Seaborn	Graphing and visualization
Pandas	Dataframe for handling dataset
Numpy	Mathematical operations
Scikit-Learn	Data preprocessing and machine learning
RDKit	Chemical processing

Students individually completed a prelaboratory worksheet before starting the experiment. This experiment is modified from the procedures detailed by Mak et al.[32] Each student was assigned either a chalcone synthesis (Figure 2, top) or a competition experiment (Figure 2, bottom). In the chalcone synthesis, the student runs a condensation reaction between one benzaldehyde and an acetophenone substrate to make an authentic product sample, whereas for the competition experiment, the student conducts the condensation between one acetophenone and two distinct benzaldehyde substrates (or two distinct acetophenone substrates and one benzaldehyde). Students running a chalcone synthesis experiment weighed or measured out (using a disposable syringe with no needle tip) their assigned benzaldehyde and acetophenone and combined these with 1 M sodium hydroxide and ethanol in a flask with a stir bar. These reagents were stirred together for ∼1 h. After stirring, students either filtered out the chalcone using a filter funnel or, if the mixture was a liquid, performed an aqueous extraction to obtain a crude mixture. They then prepared a sample of their product for gas chromatographic analysis and identified the retention time of the synthesized chalcone. Students running a competition experiment weighed or measured out (using a disposable syringe with no needle tip) their assigned benzaldehyde(s) and acetophenone(s). Importantly, they first combined the competing substrates with 1 M sodium hydroxide and ethanol in a flask with a stir bar, allowing the reagents to stir together for a few minutes before adding in the limiting reagent. After the mixture was stirred, students performed an aqueous extraction to obtain their crude mixture. They then prepared a sample of their mixture of products for gas chromatographic analysis and identified the retention times and areas of the synthesized chalcone products.

For the second half of the experiment, students worked together to extract the relevant analytical data to construct a Hammett plot in a shared Google Colab notebook. The notebook is designed for students to work with even if they have little to no prior programming experience. In this process, they completed the missing blanks throughout the notebook to familiarize themselves with the basic Python language. The notebook is divided into sections called "cells" that accomplish a certain subtask. Each cell/subtask is taught stepwise to enhance the students' understanding of the data science

https://doi.org/10.1021/acs.jchemed.5c00994
J. Chem. Educ. XXXX, XXX, XXX−XXX

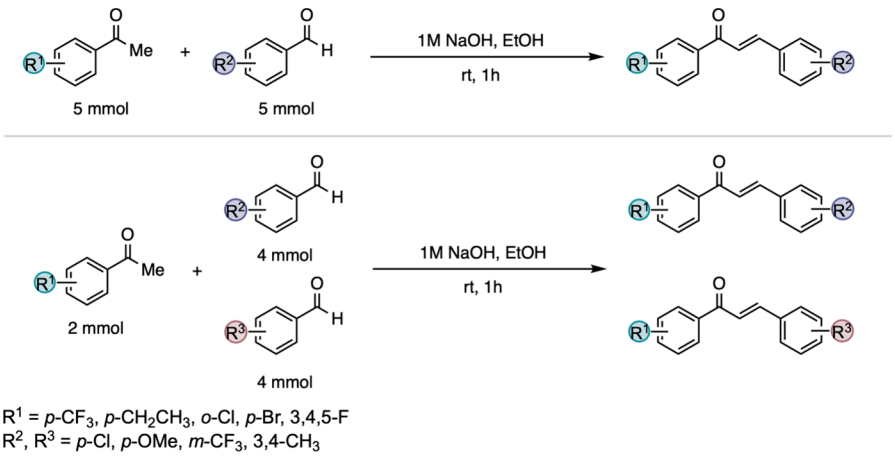


Figure 2. (top) Conditions for chalcone synthesis with substituted acetophenone and substituted benzaldehyde. (bottom) Conditions for the competition experiment with substituted acetophenone and different benzaldehydes.

pipeline while not requiring programming expertise. Students also used density functional theory (DFT)-calculated features to model the experimental data. In the Colab notebook and in a postlaboratory worksheet, students were asked to critically analyze the use of incorporating data science tools and basic machine learning algorithms to investigate the reaction. Further details of this experiment are described in the Supporting Information.

HAZARDS

At the beginning of the lab, students were given a safety briefing, including the use of personal protective equipment (PPE). Students were warned that they would be excluded if they breached the safety requirements. Closed-toed shoes, long pants/skirts covering the ankles, safety glasses, gloves, and flame-resistant laboratory coats must be worn at all times. All hazardous materials should be handled and disposed of in accordance with the recommendations of their Safety Data Sheets and Environmental Health and Safety. Sodium hydroxide is corrosive and can cause burns to skin, eyes, and respiratory tract. Ethyl acetate is flammable and a volatile organic solvent and should be used in the fume hood at all times. Benzaldehydes and acetophenones are harmful and irritating; care should be taken with the starting materials and products to avoid inhalation or contact with skin.

RESULTS AND DISCUSSION

Experimental Section

The in-laboratory portion of this experiment was modified from the procedures detailed by Mak et al.[32] The study reported the construction of standard Hammett plots to model reactivity and served as a useful template upon which to build. We were curious how chemical features obtained from DFT calculations could enhance the interpretation of substituent effects in this Claisen−Schmidt aldol condensation reaction. Therefore, we sought to include benzaldehyde and acetophenone substrates with substitution other than at the para position or at multiple locations around the benzene ring. We found that some of these substrates that we tested (see the Supporting Information) did not form solid chalcones that, as was described in the original procedure, could be filtered out of solution. For these substrate combinations, we performed a simple aqueous extraction to obtain the product using ethyl acetate as a greener and safer alternative solvent to dichloromethane. For both the chalcone synthesis and competition experiments, product purity and/or workup quality was not critical given that gas chromatography was used for analysis, in which the relevant peaks would be distinctly identified and separated from impurities by their retention times (see the Notes for Instructors for further details). Compared with the procedure by Mak et al.,[32] we increased the number of equivalents of competing substrates relative to the limiting reagent in the competition experiment such that rate data could be more accurately approximated by the obtained yield data. Given the structural similarity of the chalcone products, response factors that would typically be required for an analytical protocol using gas chromatography were omitted for experimental convenience. We implemented this experiment over two terms in which the columns used for the gas chromatography instrument differed and did not observe significant error in the subsequent data analysis.

The optimized experimental protocol was implemented during two terms of an advanced undergraduate chemistry laboratory course, in which the students were either chemistry or biochemistry majors. Students were introduced to linear free energy relationships (LFERs) and Hammett plots as well as density functional theory and machine learning in chemistry during the lecture meetings for the course prior to undertaking the laboratory experiment. After the lecture, they read the laboratory handout, which detailed the underlying concepts and purposes of the experiment as well as experimental details, and completed the prelaboratory worksheet, which was designed to ensure that they understood the safety considerations and experimental protocol.

Each student was assigned either a chalcone synthesis or a competition experiment. All 37 students across the two terms successfully carried out their experiments and made a GC sample to give to the teaching assistant for GC analysis.

After the chromatograms were obtained, at the beginning of the second part of the lab, students worked together as a class to determine the ratio of one substrate to another (k_X/k_H) in the competition experiments using the retention time data from the students who performed the product synthesis. Residual starting material and other impurities were seen in the GC chromatograms; however, these did not affect the relevant peaks. The students recorded these retention times in a shared Google sheet. After aggregation with the literature data from

https://doi.org/10.1021/acs.jchemed.5c00994
J. Chem. Educ. XXXX, XXX, XXX−XXX

Mak et al.[32] and the in-house dataset used in developing this lab, the data collected were used for analysis in the next section.

Computational Analyses

Next, students performed a series of computational and statistical analyses to investigate the experimental data. The code was provided through a partially filled Google Colab notebook written in python (example shown in Figure 3).

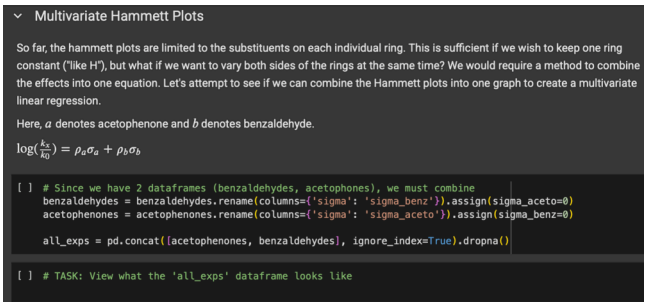


Figure 3. Screenshot of the partially filled Google Colab notebook.

Colab allows students to access and execute the code with minimal setup, irrespective of local machine capabilities. Python was chosen as the primary language for the reasons listed in Software Details (vide supra).

The students first analyzed the rate data by constructing a classical LFER to investigate the influence of substituent effects on the buildup of charge in the rate-controlling transition state of the reaction. Examples of Hammett plots varying substituents on acetophenone and benzaldehyde are shown in Figure 4. The nitro substituent (NO_2) is an outlier in this dataset, which is a common phenomenon in Hammett correlations due to field and solvation effects.[28,33] The study by Mak et al. proposed poor substrate solubility and a change in the rate-determining step to explain this specific data point.[32] This outlier was further confirmed statistically through both the interquartile-range rule and the Z-score test (see the Supporting Information for details). Note that the R^2 value[34,35] for the Hammett equation is lower than the values reported by Mak et al.[32] (0.99 for both plots with fewer samples) and the ρ value has the same sign but is of different magnitude, especially for the benzaldehyde LFER (acetophenone, 1.59; benzaldehyde, 3.09). These differences are most likely due to the increased variability in the compilation of experimental data from multiple students.

Since classical LFERs only demonstrate the effect of the substituents on one of the aryl rings in the resulting chalcone, a multivariate LFER model[36] that combines the Hammett parameters of the substituents of the acetophenones (a) and the benzaldehydes (b) was attempted.

$$\log\left(\frac{k_X}{k_H}\right) = \rho_a\sigma_a + \rho_b\sigma_b$$

The multivariate plot shown in Figure 5 showed a similar correlation with experimental data compared to both univariate Hammett plots, as shown by the R^2 of 0.90. Furthermore, the model is able to accommodate more data (both the acetophenone data and benzaldehyde data), allowing it to be more robust toward outliers, as exhibited by the R^2 of 0.65 compared to the univariate model's R^2 of 0.34 (see Figure 4). Though the multivariate LFER incorporates the impact of substituents on both reactants, the magnitudes of each ρ value are less interpretable than the ρ value arising from a univariate LFER since the two terms are not independent. Moreover, both the univariate and multivariate models cannot accommodate aryl rings with multiple substituents.

Therefore, the students were introduced to a machine learning approach in linear regression with DFT-derived features. DFT features (M06-2X/def2-TZVP) were calculated using Auto-QChem,[37] a software developed in our laboratory that calculates DFT features from SMILES strings. It provides

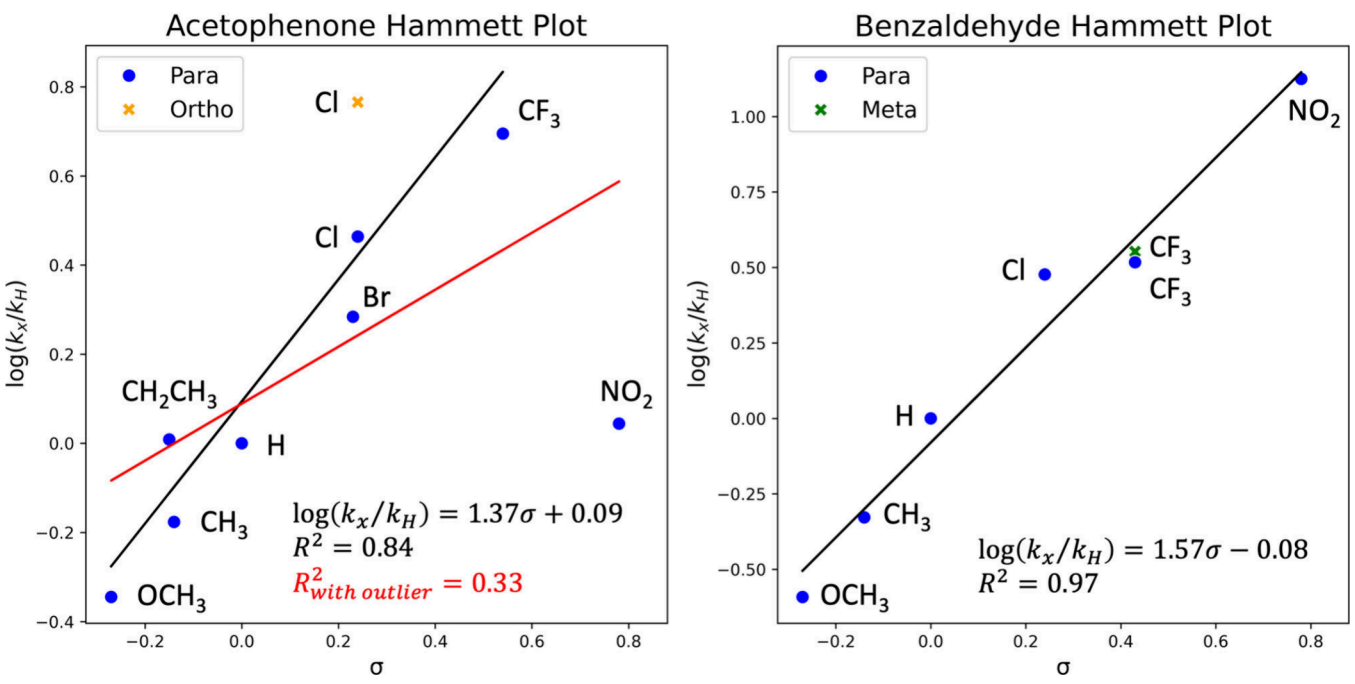


Figure 4. Hammett plots of Claisen−Schmidt condensation when substituents of acetophenone (left) and benzaldehyde (right) are varied.

https://doi.org/10.1021/acs.jchemed.5c00994
J. Chem. Educ. XXXX, XXX, XXX−XXX

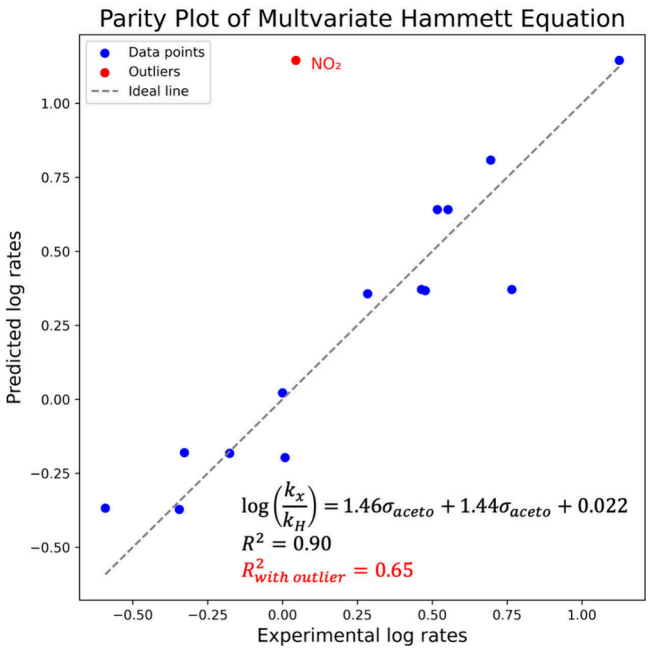


Figure 5. Parity plot of the multivariate Hammett equation.

both molecule-wide features such as dipole moment, energy, and HOMO−LUMO gap as well as atom-level features such as partial charge, electronegativity, and buried volume.[38] Though Auto-QChem streamlines the DFT calculation process, it is not necessary, and any DFT software and featurization techniques can be used.[39−41] Simultaneously, machine learning techniques were introduced to the students, including statistical and chemical preprocessing, hyperparameter optimization, and the use of ML libraries such as Scikit-learn. We believe that this approach to building a model lends itself to an apt extension from the traditional linear regression methods to which the students have likely already been exposed in introductory mathematics or statistics courses. This model was purposefully unoptimized to allow the students to experiment with the hyperparameters in order to further optimize the model. Subsequently, the class performed hyperparameter optimization, where students participated in a leaderboard challenge to build more accurate models. After optimization, the students analyzed the atomic and molecular features that were most influential in improving the accuracy of their model. Students were expected to comment on the accuracy and interpretability of this linear regression model in comparison to those of the univariate and multivariate Hammett plots. Following completion of the laboratory, the students were required to complete a postlaboratory worksheet to review key concepts from this experiment.

As can be seen in Figure 6, the multivariate Hammett plot was able to incorporate data from the Hammett studies at a high correlation (R^2 = 0.90) while also allowing inclusion of data from experiments where both arenes are simultaneously varied. According to the R^2 values, the DFT model exhibits a lower correlation (R^2 = 0.71) than the univariate and multivariate Hammett plots. However, the DFT model potentially facilitates mechanistic interpretation across a broader reaction domain. Though dependent on each student's models and each class's experimental results, in almost all of

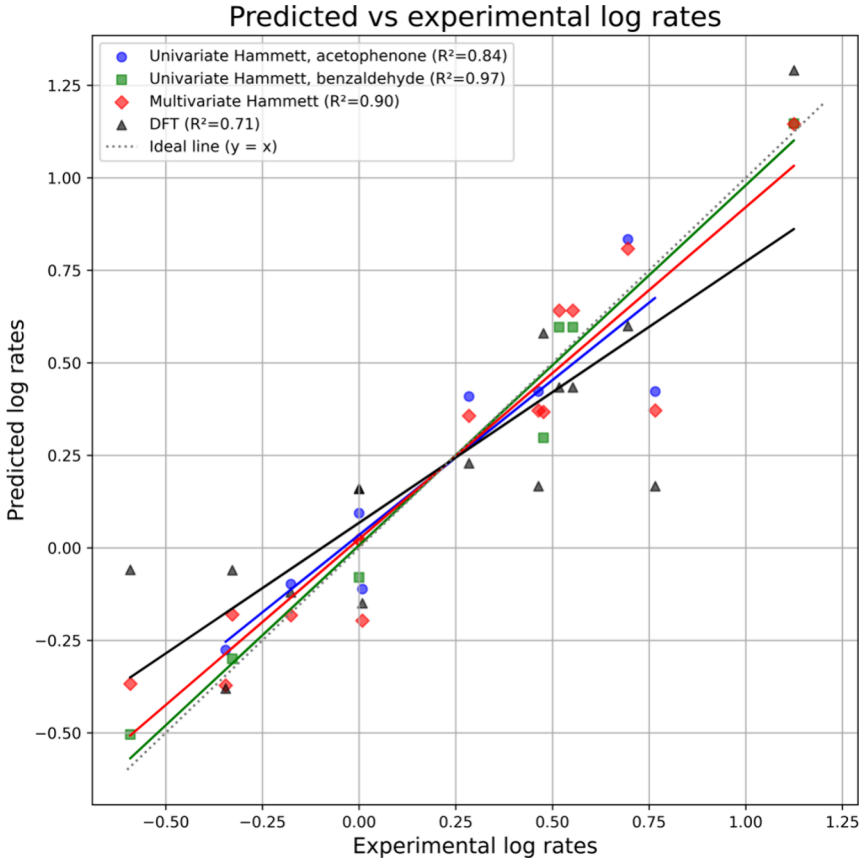


Figure 6. Predicted vs experimental rates for each model.

https://doi.org/10.1021/acs.jchemed.5c00994
J. Chem. Educ. XXXX, XXX, XXX−XXX

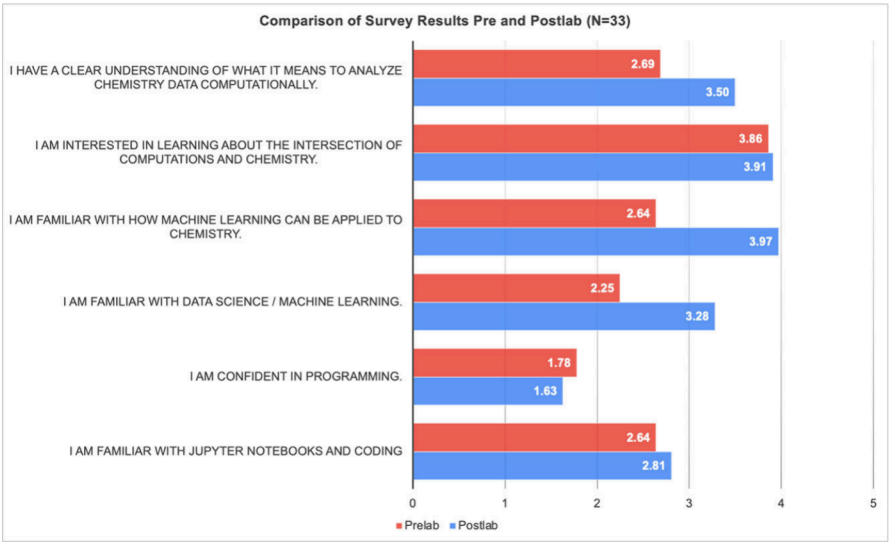


Figure 7. Bar graph comparison of survey results pre- and postmodule

the models, features centered on the carbonyl carbon of both the acetophenone and benzaldehyde were chosen as most important in constructing the DFT model. Even in the unoptimized model, the relevant features were the partial charge of the carbonyl carbon of acetophenone (NBO charge), predicted NMR shift of the carbonyl carbon of the benzaldehyde, and the LUMO energy of the benzaldehyde. This is in alignment with the mechanistic hypothesis that the nucleophilic addition step is the rate-determining step in the condensation reaction.

Learning Outcome

This experiment was implemented across two quarters at UCLA, in the first of which there were 23 students total and in the second of which there were 14 students total. Across these quarters, students earned an average of 90.7% (std. dev. = 10.4%) on their prelaboratory reports and 94.4% (std. dev. = 6.8%) on their postlaboratory reports for the lab experiment. In comparison, across the entire quarter, students earned an average of 88.4% (std. dev. = 5.3%) on their prelaboratory reports and 90.6% (std. dev. = 2.5%) on their postlaboratory reports. These results demonstrate excellent comprehension of the material following the lectures, and this two-part experiment is designed to promote experiential learning. The prelaboratory questions were structured to connect core chemistry principles with machine learning, prompting students to apply familiar concepts in a new analytical context. This experiment focused on chemistry concepts that could be applied to a machine learning context, allowing students to practice higher-order thinking by investigating and thinking critically about general concepts of machine learning in chemical contexts. Meanwhile, the postlaboratory questions involved analysis of the machine learning model they obtained while running the code, such as graph behavior, data analysis, and potential improvements to the model.

We collected students' feedback on the lab through a guided reflection question on their postlaboratory report. Students from the first quarter in which this experiment was implemented commented on the straightforwardness and ease of the experimental wet-lab portion and appreciated the opportunity to collaborate with their peers in building a dataset. They thought that the data science portion of the lab was interesting and relevant to their other studies. There were mixed comments on the difficulty of the data analysis, with suggestions for further explanation of the analysis process and code. Thus, for the next quarter in which this experiment was implemented, we modified the code to be more interactive for students, providing opportunities for them to fill out the functions. The feedback received was similar to the first quarter and continued to be mixed, the major suggestion of which was to increase the explanation for the data analysis section of the lab. We discuss later in this section possible improvements to this lab to address this student feedback.

To better quantify the students' attitudes toward the incorporation of machine learning in this hybrid experimental and data analysis context, we asked the students to fill out voluntary anonymous surveys before and after the module. Of the statements, six were present in both surveys:

- Statement 1: I have a clear understanding of what it means to analyze chemistry data computationally.
- Statement 2: I am interested in learning about the intersection of computations and chemistry.
- Statement 3: I am familiar with how machine learning can be applied to chemistry.
- Statement 4: I am familiar with data science/machine learning.
- Statement 5: I am confident in programming.
- Statement 6: I am familiar with Jupyter notebooks/Google Colab and coding.

This allowed for a quantitative comparison, as shown in Figure 7. The survey was completed by 33 of the 37 students (89%). Most notably, statements regarding the student's familiarity and exposure to how computational tools can be applied to chemistry (statements 1 and 3) saw notable increases postlab (30.1% and 50.4%, respectively). Students showed a substantial increase (45.8%) in their familiarity with data science and machine learning, even when it is not in the context of chemistry. Students' familiarity with Jupyter notebooks/Google Colabs and coding in general also increased after the lab, though to a lower extent, as shown in statement 6. Statement 2 further shows that the students are still interested in learning more about the intersection of computational tools and chemistry.

https://doi.org/10.1021/acs.jchemed.5c00994
J. Chem. Educ. XXXX, XXX, XXX−XXX

We observed an 8.4% decrease in the students' confidence in programming, as seen by statement 5. This may be attributed to the fact that the code was written to accommodate students with no prior programming experience, often abstracting and skipping details of the code during the discussion. We believe this can be alleviated by either having longer modules where more detailed discussion of the code can be taught or requiring an Introduction to Python course.

SUMMARY

We developed a two-part laboratory module that introduces undergraduate chemistry students to modern computational tools for data analysis and interpretation of substituent effects and reactivity in a classic organic reaction. Both the synthesis and competition experiments in the experimental section are straightforward to carry out in the laboratory, enabling students to practice essential organic chemical reaction setup and workup techniques. In the second module, students collaborate to assemble a shared dataset which they use to train machine learning models using physical organic principles, statistical methods, and computational chemistry tools. Given the current directions of contemporary research, we envision that this experiment will be useful to introduce into the undergraduate chemistry curricula, connecting students at an earlier stage to potential future research pathways.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available at https://pubs.acs.org/doi/10.1021/acs.jchemed.5c00994.

General information, characterization of products, GC data, Google Colab notebooks, notes for instructors, and survey questions (PDF)

Student handout and key (PDF)

Blank lecture notes (PDF)

Filled lecture notes (PDF)

AUTHOR INFORMATION

Corresponding Author

Abigail G. Doyle − *Department of Chemistry and Biochemistry, University of California, Los Angeles, California 90095, United States;* orcid.org/0000-0002-6641-0833; Email: abigaildoyle@ucla.edu

Authors

Daniel S. Min − *Department of Chemistry and Biochemistry, University of California, Los Angeles, California 90095, United States*

Flora Fan − *Department of Chemistry and Biochemistry, University of California, Los Angeles, California 90095, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jchemed.5c00994

Author Contributions

‡D.S.M. and F.F. contributed equally.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank the students of the Chem 30CL organic chemistry laboratory course (Fall 2024 and Winter 2025) for their feedback on this experiment. We thank Dr. Nicholas Deifel and Dr. Soumitra Athavale for supporting the incorporation of this experiment in their syllabus and the teaching assistants (Vicki Rubio, Ethan Prout, Calvin Ho, James Collings, Andrew Baublis, Haoding Lin, and Daniel Pan) for their help in implementing this experiment. We thank Natalie Ha and Maria Dimaano-Salanga for procuring the reagents and setting up the materials required for this experiment. We thank Hannah Whang Sayson for her help in constructing survey questions and Prof. Arlene Russell for her help in obtaining IRB support (IRB-25-0936). These studies were supported by project (IIP #23-01) funding from the UCLA Committee on Instructional Improvement Programs (CIIP), the United States National Science Foundation (NSF) Office of Advanced Cyberinfrastructure (OAC-2118201), and shared instrumentation grants from the National Science Foundation under equipment grants CHE-1048804 and CHE-2117480, along with the NIH Office of Research Infrastructure Program Supergrant S10OD028644.

REFERENCES

(1) Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Hunter, C. A.; Bekas, C.; Lee, A. A. Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *ACS Cent. Sci.* **2019**, *5* (9), 1572−1583.

(2) Żurański, A. M.; Martinez Alvarado, J. I.; Shields, B. J.; Doyle, A. G. Predicting Reaction Yields via Supervised Learning. *Acc. Chem. Res.* **2021**, *54* (8), 1856−1865.

(3) Segler, M. H. S.; Waller, M. P. Neural-Symbolic Machine Learning for Retrosynthesis and Reaction Prediction. *Chem.—Eur. J.* **2017**, *23* (25), 5966−5971.

(4) Coley, C. W.; Barzilay, R.; Jaakkola, T. S.; Green, W. H.; Jensen, K. F. Prediction of Organic Reaction Outcomes Using Machine Learning. *ACS Cent. Sci.* **2017**, *3* (5), 434−443.

(5) Torres, J. A. G.; Lau, S. H.; Anchuri, P.; Stevens, J. M.; Tabora, J. E.; Li, J.; Borovika, A.; Adams, R. P.; Doyle, A. G. A Multi-Objective Active Learning Platform and Web App for Reaction Optimization. *J. Am. Chem. Soc.* **2022**, *144* (43), 19999−20007.

(6) Wang, J. Y.; Stevens, J. M.; Kariofillis, S. K.; Tom, M.-J.; Golden, D. L.; Li, J.; Tabora, J. E.; Parasram, M.; Shields, B. J.; Primer, D. N.; Hao, B.; Del Valle, D.; DiSomma, S.; Furman, A.; Zipp, G. G.; Melnikov, S.; Paulson, J.; Doyle, A. G. Identifying General Reaction Conditions by Bandit Optimization. *Nature* **2024**, *626* (8001), 1025−1033.

(7) Coley, C. W.; Thomas, D. A., III; Lummiss, J. A. M.; Jaworski, J. N.; Breen, C. P.; Schultz, V.; Hart, T.; Fishman, J. S.; Rogers, L.; Gao, H.; Hicklin, R. W.; Plehiers, P. P.; Byington, J.; Piotti, J. S.; Green, W. H.; Hart, A. J.; Jamison, T. F.; Jensen, K. F. A Robotic Platform for Flow Synthesis of Organic Compounds Informed by AI Planning. *Science* **2019**, *365* (6453), No. eaax1566.

(8) Zhong, Z.; Song, J.; Feng, Z.; Liu, T.; Jia, L.; Yao, S.; Hou, T.; Song, M. Recent Advances in Deep Learning for Retrosynthesis. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2024**, *14* (1), No. e1694.

(9) Call, A.; Palone, A.; Liles, J. P.; Romer, N. P.; Read, J. A.; Luis, J. M.; Sigman, M. S.; Bietti, M.; Costas, M. Understanding Catalytic Enantioselective C−H Bond Oxidation at Nonactivated Methylenes Through Predictive Statistical Modeling Analysis. *ACS Catal.* **2025**, *15* (3), 2110−2123.

(10) Kim, S. F.; Liles, J. P.; Lux, M. C.; Park, H.; Jurczyk, J.; Soda, Y.; Yeung, C. S.; Sigman, M. S.; Sarpong, R. Interrogation of Enantioselectivity in the Photomediated Ring Contractions of Saturated Heterocycles. *J. Am. Chem. Soc.* **2025**, *147* (2), 1851−1866.

(11) Romer, N. P.; Min, D. S.; Wang, J. Y.; Walroth, R. C.; Mack, K. A.; Sirois, L. E.; Gosselin, F.; Zell, D.; Doyle, A. G.; Sigman, M. S.

Data Science Guided Multiobjective Optimization of a Stereoconvergent Nickel-Catalyzed Reduction of Enol Tosylates to Access Trisubstituted Alkenes. *ACS Catal.* **2024**, *14* (7), 4699−4708.
(12) Baum, Z. J.; Yu, X.; Ayala, P. Y.; Zhao, Y.; Watkins, S. P.; Zhou, Q. Artificial Intelligence in Chemistry: Current Trends and Future Directions. *J. Chem. Inf. Model.* **2021**, *61* (7), 3197−3212.
(13) Santiago, C. B.; Guo, J.-Y.; Sigman, M. S. Predictive and Mechanistic Multivariate Linear Regression Models for Reaction Development. *Chem. Sci.* **2018**, *9* (9), 2398−2412.
(14) Haas, B. C.; Goetz, A. E.; Bahamonde, A.; McWilliams, J. C.; Sigman, M. S. Predicting Relative Efficiency of Amide Bond Formation Using Multivariate Linear Regression. *Proc. Natl. Acad. Sci. U.S.A.* **2022**, *119* (16), No. e2118451119.
(15) Crawford, J. M.; Kingston, C.; Toste, F. D.; Sigman, M. S. Data Science Meets Physical Organic Chemistry. *Acc. Chem. Res.* **2021**, *54* (16), 3136−3148.
(16) Tran, R.; Brunskill, V.; Musgrove, A.; Sutherland, T. C.; Derksen, D. J. A Problem-Based Introduction to Machine Learning in the Undergraduate Organic Chemistry Laboratory: Prediction of Diels−Alder Reaction Rates. *J. Chem. Educ.* **2025**, *102* (8), 3443−3451.
(17) Thrall, E. S.; Martinez Lopez, F.; Egg, T. J.; Lee, S. E.; Schrier, J.; Zhao, Y. Rediscovering the Particle-in-a-Box: Machine Learning Regression Analysis for Hypothesis Generation in Physical Chemistry Lab. *J. Chem. Educ.* **2023**, *100* (12), 4933−4940.
(18) Jiang, S.; McClure, J.; Mao, H.; Chen, J.; Liu, Y.; Zhang, Y. Integrating Machine Learning and Color Chemistry: Developing a High-School Curriculum toward Real-World Problem-Solving. *J. Chem. Educ.* **2024**, *101* (2), 675−681.
(19) Lafuente, D.; Cohen, B.; Fiorini, G.; García, A. A.; Bringas, M.; Morzan, E.; Onna, D. A Gentle Introduction to Machine Learning for Chemists: An Undergraduate Workshop Using Python Notebooks for Visualization, Data Processing, Analysis, and Modeling. *J. Chem. Educ.* **2021**, *98* (9), 2892−2898.
(20) Ziegler, B. E. Theoretical Hammett Plot for the Gas-Phase Ionization of Benzoic Acid versus Phenol: A Computational Chemistry Lab Exercise. *J. Chem. Educ.* **2013**, *90* (5), 665−668.
(21) Rainey, M. A.; Benda, M. C.; Mayberry, K. A.; Smeekens, J. M.; Braga, R. A.; Bottomley, L. A.; O'Mahony, C. M. Data Science Meets Mineral Analysis: An Innovative Laser-Induced Breakdown Spectroscopy Experiment for Undergraduate Chemistry Students. *J. Chem. Educ.* **2024**, *101* (7), 2869−2879.
(22) Mahjour, B.; McGrath, A.; Outlaw, A.; Zhao, R.; Zhang, C.; Cernak, T. Interactive Python Notebook Modules for Chemoinformatics in Medicinal Chemistry. *J. Chem. Educ.* **2023**, *100* (12), 4895−4902.
(23) Kearnes, S. M.; Maser, M. R.; Wleklinski, M.; Kast, A.; Doyle, A. G.; Dreher, S. D.; Hawkins, J. M.; Jensen, K. F.; Coley, C. W. The Open Reaction Database. *J. Am. Chem. Soc.* **2021**, *143* (45), 18820−18826.
(24) Rana, D.; Pflüger, P. M.; Hölter, N. P.; Tan, G.; Glorius, F. Standardizing Substrate Selection: A Strategy toward Unbiased Evaluation of Reaction Generality. *ACS Cent. Sci.* **2024**, *10* (4), 899−906.
(25) Kolb, D. A. *Experiential Learning: Experience as the Source of Learning and Development*; Prentice Hall: Englewood Cliffs, NJ, 1984.
(26) Anderson, L.; Krathwohl, D. A. *Taxonomy for Learning, Teaching and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*; Longman: New York, 2001.
(27) Bloom, B. S.; Engelhart, M. D.; Furst, E. J.; Hill, W. H.; Krathwohl, D. R. *Taxonomy of Educational Objectives, Handbook I: Cognitive Domain*; David McKay Company: New York, 1956.
(28) Hansch, C.; Leo, A.; Taft, R. W. A Survey of Hammett Substituent Constants and Resonance and Field Parameters. *Chem. Rev.* **1991**, *91* (2), 165−195.
(29) Hammett, L. P. Some Relations between Reaction Rates and Equilibrium Constants. *Chem. Rev.* **1935**, *17* (1), 125−136.
(30) Berardi, M. D.; Gentile, F.; Kozik, I.; Gregg, T. M. Aldol Condensation Reaction Rate Demonstrates Steric and Electronic Substituent Effects in the Organic Chemistry Lab. *J. Chem. Educ.* **2021**, *98* (5), 1732−1735.
(31) Bain, R. M.; Pulliam, C. J.; Yan, X.; Moore, K. F.; Müller, T.; Cooks, R. G. Mass Spectrometry in Organic Synthesis: Claisen−Schmidt Base-Catalyzed Condensation and Hammett Correlation of Substituent Effects. *J. Chem. Educ.* **2014**, *91* (11), 1985−1989.
(32) Mak, K. K. W.; Chan, W.-F.; Lung, K.-Y.; Lam, W.-Y.; Ng, W.-C.; Lee, S.-F. Probing the Rate-Determining Step of the Claisen−Schmidt Condensation by Competition Reactions. *J. Chem. Educ.* **2007**, *84* (11), 1819.
(33) Sessa, F.; Olsson, M.; Söderberg, F.; Wang, F.; Rahm, M. Experimental Quantum Chemistry: A Hammett-inspired Fingerprinting of Substituent Effects. *ChemPhysChem* **2021**, *22* (6), 569−576.
(34) R^2 was used because students are already exposed to this metric in earlier coursework and can critically assess its limitations as a potential source of error. It is noted that although commonly applied, R^2 can be misleading since it primarily measures how well the data align along a straight line, regardless of its position; a line with incorrect slope or intercept can still yield a high R^2, masking systematic bias or poor predictive accuracy. Root-mean-square error or mean absolute error can provide more reliable measures of model performance. See: Kvalseth, T. O. Cautionary Note about R^2. *Am. Stat.* **1985**, *39*, 279−285.
(35) Aboal-Somoza, M.; Crujeiras, R. M. Misuse of Linear Regression Technique in Analytical Chemistry? *J. Chem. Educ.* **2024**, *101* (3), 1062−1070.
(36) Endo, S. Applicability Domain of Polyparameter Linear Free Energy Relationship Models Evaluated by Leverage and Prediction Interval Calculation. *Environ. Sci. Technol.* **2022**, *56* (9), 5572−5579.
(37) Żurański, A. M.; Wang, J. Y.; Shields, B. J.; Doyle, A. G. Auto-QChem: An Automated Workflow for the Generation and Storage of DFT Calculations for Organic Molecules. *React. Chem. Eng.* **2022**, *7* (6), 1276−1284.
(38) Clavier, H.; Nolan, S. P. Percent Buried Volume for Phosphine and N-Heterocyclic Carbene Ligands: Steric Properties in Organometallic Chemistry. *Chem. Commun.* **2010**, *46* (6), 841−861.
(39) Antle, J. P.; Kimura, M. W.; Racioppi, S.; Damon, C.; Lang, M.; Gatley-Montross, C.; Sánchez B., L. S.; Miller, D. P.; Zurek, E.; Brown, A. M.; Gast, K.; Simpson, S. M. Applying Density Functional Theory to Common Organic Mechanisms: A Computational Exercise. *J. Chem. Educ.* **2023**, *100* (1), 355−360.
(40) Rowley, C. N.; Woo, T. K.; Mosey, N. J. A Computational Experiment of the Endo versus Exo Preference in a Diels−Alder Reaction. *J. Chem. Educ.* **2009**, *86* (2), 199.
(41) Hirschi, J. S.; Bashirova, D.; Zuehlsdorff, T. J. Opening the Density Functional Theory Black Box: A Collection of Pedagogic Jupyter Notebooks. *J. Chem. Educ.* **2023**, *100* (11), 4496−4503.

https://doi.org/10.1021/acs.jchemed.5c00994
J. Chem. Educ. XXXX, XXX, XXX−XXX