ONLINE GRADUATE CERTIFICATE IN DATA SCIENCE FOR THE CHEMICAL INDUSTRY

Andrew J. Medford, Fani Boukouvala, Martha A. Grover, David Sholl, Carson Meredith, Pengfei Cheng, Sihoon Choi, Gabriel S. Gusmão, Zachary Kilwein, Suryateja Ravutla, Fatimah Wirth, Jennifer Wooley, and Zaid Sewer
Georgia Institute of Technology • Atlanta, GA 30332

INTRODUCTION

In the past decade, society has been transformed by a revolution in the collection and use of data. Over roughly the same period, the chemical process industry (CPI) has gone through a manufacturing renaissance in the US, with hundreds of billion dollars of construction on the Gulf Coast alone. The CPI employs more than 800,000 people in the US and has created products worth more than $700 billion in 2017.[1] Despite the vast scale of the CPI, which encompasses oil and gas, chemicals, consumer products, and pharmaceuticals, this sector lags behind other areas in taking advantage of data science.[2] The CPI is an interdisciplinary field that employs chemical engineers, chemists, material science engineers and bioengineers, all of whom need to design, operate and optimize materials, formulations, processes, and products. A 2018 McKinsey study estimated that incorporating artificial intelligence in operations within the global CPI would realize $800 billion in annual value.[2] However, according to industry feedback, simply embedding data scientists without CPI training in risky chemical manufacturing environments does not lead to meaningful change in operations, and oftentimes leads to distrust of data-science from domain experts.[3] At the same time, the CPI's current employees have domain expertise and first-hand experience of the effects of digitalization, but often lack the necessary programming and data science background. A key need in the US CPI is development of a highly skilled workforce that combines domain expertise in industrial-scale chemicals processing with robust and adaptable data science skills.[2,4-8]

In response to this need, a new program was created at the Georgia Institute of Technology (GT), in the form of an online graduate certificate. The central aim of the "Online Graduate Certificate in Data Science for the Chemical Industry" (DSCI) program is to *create the opportunity for current undergraduates, graduate students, and CPI professionals to be trained together, to become part of the next generation of citizen data scientists that the CPI needs*.[9] Industrial practitioners who complete this training are equipped to lead data-driven transformations in their companies and aim to see their own personal career prospects grow. Current Georgia Tech students who complete the program interact with CPI professionals with practical and business experience in the context of using real industrial data.

Graduate Certificates or "mini-Masters" degrees at Georgia Tech are a unique model for graduate education that is available to both degree and non-degree students and require 12 credit hours for completion. We have developed a *fully online* graduate certificate that involves two core courses designed and offered by chemical engineering faculty focused on foundations of data science in an application-rich context. These are followed by two courses chosen from a range of electives already offered online across campus on topics of data analytics (or, equivalently, data science) and machine learning (ML). Motivated by similar goals, other programs have been created in recent years. The MS Degree offered in the University of Washington was one of the first to offer a data-analytics degree specific to the chemical sciences, with a comprehensive program of 39 credits.[10] Other full MS degrees offered by chemical engineering departments include the MS in Chemical Engineering with a Concentration in Data and Computational Science at Columbia University, Purdue University's Data-Science in Chemical Engineering MS, and Cornell University's Computational Informatics MS.[11-13] Finally, the "Graduate Data Science Certificate Program" offered at the University of Michigan is the another offering with a similar number of credits as our program, although it is not specific to chemical engineering.[14] The program offered at Georgia Tech is different than the above

initiatives because it is designed to be a short program (mini-MS with the possibility of extending to a full MS degree), and it is offered in an asynchronous online format designed for professionals.

Extensive discussions with industry colleagues indicated that our proposed program fills a critical gap in workforce development for the CPI. Although generic data science "boot camps" and MS degrees in data analytics exist, these resources do not provide sufficient depth on the challenges that are relevant to the data-driven applications emerging in the CPI. Specifically, they do not put enough focus on the objectives and constraints that are critical to the CPI, such as first-principles (i.e., mass and energy balances, thermodynamics, kinetics, etc.), safety and engineering ethics, impact on energy efficiency and human health. The lack of domain specific focus and applications in these "generic" data science settings might limit the ability of individuals trained by them to readily translate these skills into the specific needs of the CPI. Further, many existing programs specific to the chemical industry take too long for many people already in the workforce who want to broaden their skillset. Finally, existing graduate programs taught by computer scientists typically assume a prior expertise in programming that may not be consistent with some CPI professionals interested on this topic.

In this paper we describe the content and teaching innovations we have incorporated in the two core courses, which aim to introduce chemical engineers to Python® programming, data processing, ML, and optimization, using specific examples relevant to the CPI. The unique elements of the content of the courses that we have designed include:

- introduction to Python programming assuming prior MATLAB®-based training
- introduction to basic statistics and ML using chemical engineering data sets (including benchmark data sets obtained by industry)
- embedding of ML models within chemical process optimization formulations
- integration of data-driven models with chemical engineering principles (physics-informed ML or hybrid modeling)

Teaching techniques that we have employed for effective online training include:

- collaboration with Instructional Designers at Georgia Tech Professional Education (GTPE) for generation of carefully curated module-based video lectures
- design of frequent online coding-based skill-check assignments using Jupyter Notebooks®, Vocareum®, and both auto-grading and peer-grading techniques
- generation of vertically integrated groups that work together throughout the semester on a project for data-driven decision making

As the first cohort of our students has graduated and the second one is underway, we will describe lessons learned and provide testimonials from students so far.

PROGRAM STRUCTURE

The DSCI program requires students to take the first core course on "Data Analytics for Chemical Engineers" (DACE) and subsequently the second core course on "Data-Driven Process Systems Engineering" (DDPSE). These two core courses have been designed and taught by chemical engineering faculty. Upon completion of these first two courses, students select two additional courses on more advanced data analytics, ML, or cybersecurity topics offered by the Computer Science and Industrial & Systems Engineering Schools at Georgia Tech. A list of potential electives includes Computing for Data Analysis, Big-Data Systems & Analytics, Database Systems Concepts & Design, Temporal, Spatial & Adaptive Databases, Data & Visual Analytics, Design & Analysis of Experiments, Stochastic Optimization, Data Mining & Statistical Learning, Information Security Policies & Strategies, and many more, for a total of 28 potential courses. In this paper we will not describe the electives in detail but will focus on the introductory core DACE and DDPSE courses. We note that all lecture materials, along with a publicly sharable dataset generously provided by Dow Chemical, are freely available via the following two GitHub® repositories: https://github.com/medford-group/data_analytics_ChE, and https://github.com/DDPSE/GT_DataDrivenPSE_Course.

Our first DSCI course offers a brief introduction to Python, but we emphasize to prospective students that to be successful in the program, they need to have prior programming experience in any language, and preferably experience with Python. If they do not have prior Python experience, the syllabus states that extra effort will be required to increase their level of expertise in Python throughout the semester. This has been a successful model that aims to bring students of different programming skills into common ground and avoids spending significant course time on basic programming concepts that can be learned through other resources. Throughout the duration of the first two core courses, students are taught (through lectures, coding-based skill-checks, and homework) advanced topics in Python-based ML and integration of ML and optimization, which ensures that the material is still challenging even for students who come in with a Python background. Moreover, the semester-long projects in each course provide an opportunity for students to define a problem that is interesting and engag-

ing at their current skill level. Looking into the future, we hypothesize that students will enter the program with more Python programming expertise. However, we expect that the omission of basic Python programming from the course, along with its "project centric" focus, will keep the material relevant and challenging into the future.

The DSCI certificate was originally designed for students external to GT, mainly industrial professionals interested in learning data analytics. However, a survey within our PhD program indicated that 75% of our graduate students were interested in completing the certificate and obtaining the degree as well as satisfying their minor requirement. Moreover, modified versions of the two core courses have been offered as electives to our undergraduate students, with increasing popularity. As a result, we embraced this unique opportunity to bring together these three populations of students and designed our core courses and assignments to take advantage of this diversity to enhance learning. The enrollment of undergraduate, graduate, and professional students in the two core courses for the first two years of course offerings is shown in Figure 1. Below, we will describe the content and learning outcomes of the DACE and DDPSE courses, followed by a description of innovations on the teaching techniques and tools used in those.

Data Analytics for Chemical Engineers Course

Chemical engineers typically receive some basic training in numerical methods and programming, but these skills are not always reinforced throughout the curriculum, and industry professionals may have not practiced programming or linear algebra in many years. In addition, related disciplines such as chemistry and materials science often have even less mathematical training in their curricula. Moreover, the recent advances in data analytics and machine learning have resulted in new techniques and new jargon that are unfamiliar to industry professionals and even current students in chemistry and chemical engineering. Although there are a plethora of introductory courses on data analytics and machine learning available, they are typically not tailored to be accessible to students with a background in chemistry and chemical engineering, and they often rely on examples that are disconnected from the chemical industry.

The DACE course seeks to fill this training gap by providing assignments and examples designed to orient students from chemical engineering and related disciplines with the fields of data analytics and machine learning. At the end of the course, students should be able to (a) apply Python libraries to analyze, visualize, and organize data, (b) select appropriate analytics models and evaluate them using quantitative metrics, (c) optimize hyperparameters and features of analytics models to improve performance. The course is organized into six key topics:

1. Numerical Methods (programming and linear algebra review)
2. Machine Learning for Regression (kernel ridgeregression, hyperparameter optimization, regression for high-dimensional data)
3. Machine Learning for Classification (generalized linear models, k-nearest neighbors, decision trees)
4. Data Management (basic data structuring, data access via API's)
5. Exploratory Data Analysis (dimensional reduction, clustering, and generative models)
6. Feature Engineering (supervised dimensional reduction, symbolic regression, time series analysis) (Figure 2)

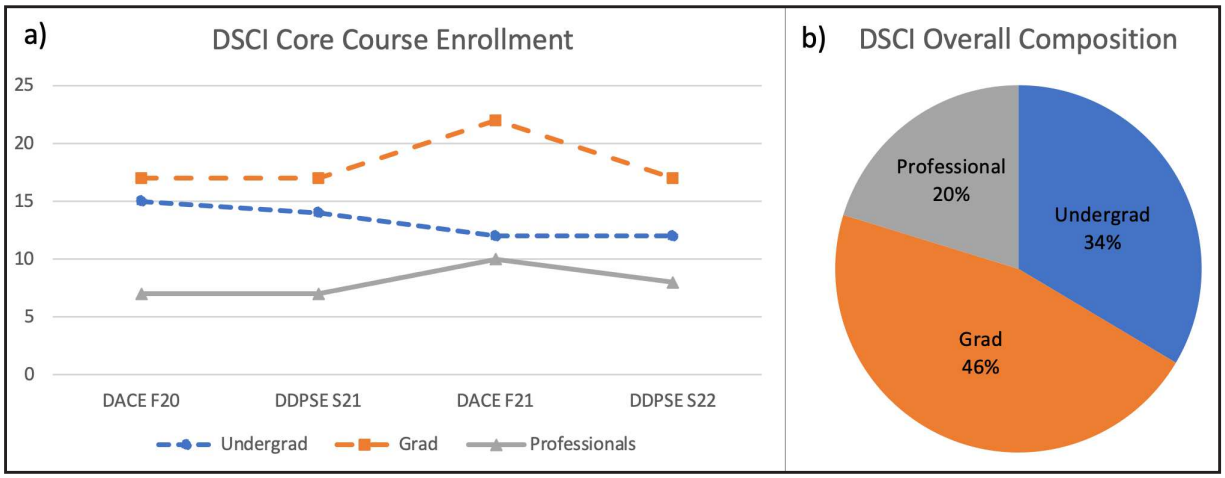

Figure 1. *Enrollment of undergraduate, graduate, and professional students in the DSCI core courses with (a) total enrollment in each course per semester and (b) overall composition averaged across all courses.*

The course starts with a quick Python primer for students familiar with MATLAB or other languages and utilizes basic exercises in numerical methods, such as orthogonalizing matrices and solving multi-dimensional non-linear optimization problems to simultaneously build competency in Python and review the basic mathematical foundations needed for data analytics and machine learning. This module is often the most difficult for students who are unfamiliar with Python or have not practiced programming in a long time, so additional office hours and review sessions are provided in the first few weeks. The subsequent modules focus on supervised models (classification and regression), with an emphasis on cross validation and hyperparameter optimization. After a midterm exam based on hands-on coding, the data management module "zooms out" and discusses common tools and best practices for how to prepare and organize data for analysis with Python or other programmatic approaches. The subsequent modules introduce unsupervised methods and feature engineering, and a coding-based comprehensive final exam assesses mastery of all topics.

Datasets in the course are chosen to be familiar to chemical engineers and include fitting peaks of spectra, predicting the synthesizability of materials, and correlating process inputs and outputs from a real chemical process dataset provided by Dow Chemical. Throughout the course students also work on a semester-long project for application of data science techniques to real datasets. These student-defined projects allow industry professionals to bring in problems from their work, or graduate students to bring in problems from their research. Students work with "vertically integrated" groups of industry professionals, graduate, and undergraduate students and work directly with instructors to define goals and build and optimize models throughout the semester. The projects are often continued in subsequent semesters in the DDPSE course.

Data-Driven Process Systems Engineering Course

Optimization problems can be found everywhere in engineering, in both industry and academia. However, most chemical engineering departments do not offer a course that discusses optimization specifically in the context of chemical engineering. At the same time, computational optimization is very tightly linked to programming, data analytics, and machine learning. In academia, data analytics coupled with optimization is being used to enable or expedite scientific discovery in materials science, pharmaceuticals, process systems engineering, and more.[3, 8, 15-20] Similarly, industry is entering an era of digitalization that has led to an explosion of chemical, process, manufacturing, and operations data.[3, 5, 6, 8, 21] Undoubtedly a large fraction of chemical engineering graduates will face design, control, or operations optimization problems that will also involve handling of data sets.

As data analytics becomes an influential tool for decision making in industry and academia, incorporation of such concepts in our curriculum will make our graduates competitive in today's market. The learning objectives for the DDPSE course are for students to be able to:

- explain the basic theory of linear, nonlinear and mixed-integer optimization
- apply sampling, regression, validation and data-reduction techniques
- use data-driven techniques for optimization
- assess the dangers and ethics of the use of data for decision making

The course material was divided into five parts:

1. Basics of Optimization Theory (formulations, linear programming, nonlinear programming, mixed-integer programming)
2. Building Data-Driven Models To Represent Constraints and Objective Functions in Optimization (data pre-processing, regression, dimensionality-reduction)
3. Data-Driven Optimization (sample-based optimization, using regression for optimization, evolutionary optimization methods)
4. Inference of Results (assessing optimality, model validation, adaptive sampling)
5. Advanced Topics (integration of first principles with data-driven regression and handling uncertainty in optimization)

Several topics introduced in the DACE course (assesement and validation, modeling, programming, and data anaylsis), are discussed again in the DDPSE course; however, the overlap is not significant because they are discussed with a different lens, specifically in the context of optimization. The interaction and overlap between the DACE and DDPSE courses are shown schematically in Figure 2. Student evaluations have indicated that the level of overlap and the reintroduction of similar topics from a different perspective are welcome.

The DDPSC course first introduces students to the basics of optimization, such as how to formulate an optimization problem by introducing the correct variables, forming the appropriate objective function and constraints, and finally, identifying the characteristics of the optimization problem (i.e., linear, nonlinear, nonconvex, mixed integer). An overview of basic optimization theory for linear, nonlinear and mixed integer optimization is also covered, along with an overview of available state-of-the-art solvers and their capabilities and limitations. Subsequently, the course discusses the challenges of dealing with data when it is not easy or

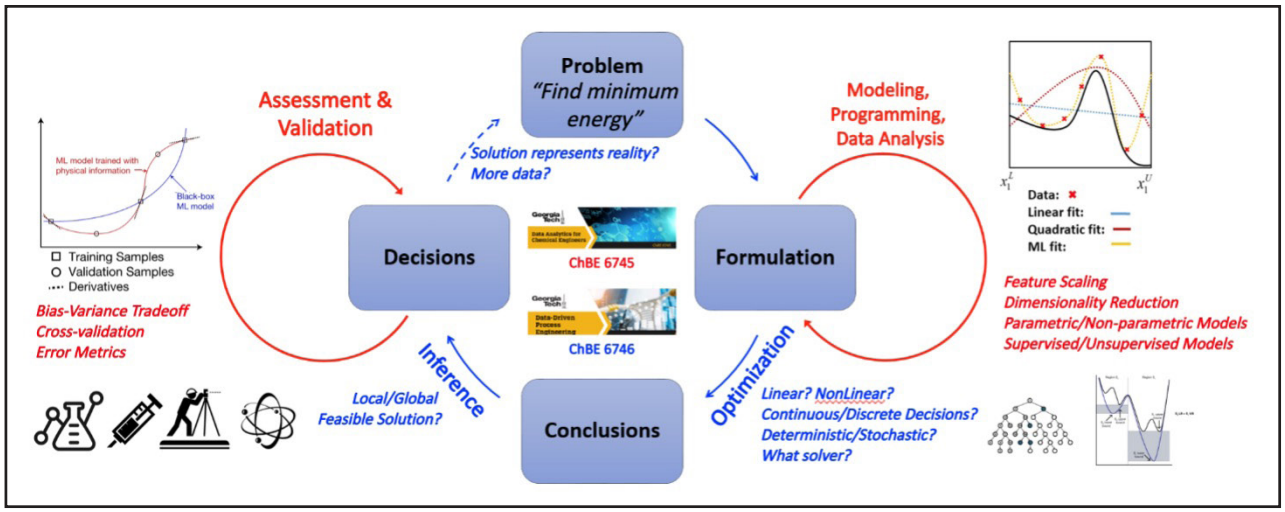

Figure 2. *Overview of the Data Analytics for Chemical Engineers (red text) and Data-Driven Process Systems Engineering courses (blue text). Assessment and validation, modeling, programming, and data analysis are first introduced in Data Analytics for Chemical Engineers, but are reviewed and placed into the broader context of optimization and problem solving in Data-Driven Process Systems Engineering.*

possible to derive the equations of the optimization formulation. In this second part of the course, we discuss how to collect data, assess the quality of a data set, and use this data for regression. Building on the DACE course, various regression, data-preprocessing, and dimensionality reduction techniques are discussed (e.g., linear and quadratic regression, generalized linear regression, neural networks, support vector regression, Gaussian processes and principal component analysis). The importance of validation, regularization, and constraining of ML models using first-principles information, as well as the danger of overfitting, are stressed, always in the context of how these affect optimization solutions. Finally, the course discusses data-driven optimization techniques. First, direct-search and trust-region algorithms are discussed. Next, the course focuses on how to embed fitted regression models within optimization formulations to locate optimal solutions. Here, the importance of model mismatch and its effects on optimality, validation of the obtained solutions, adaptive sampling, and optimal design of experiments are discussed. The course concludes with a final advanced topics module discussing uncertainty caused by data and how this can be handled in optimization using stochastic programming or robust optimization.

The course introduces students to Python and Pyomo®,[22] while the majority of the lectures are in the form of Jupyter Notebooks.[23] This platform enables interactive problem solving and data visualization. The largest component of the course grade is based on a semester-long group project. Students identify, formulate and solve their own data-driven optimization problem, either by using their research data or by using publicly available databases.

Teaching Innovations

Module-Based Lecture Videos. In both courses, the instructors worked with GTPE to record curated, high-quality short videos that were released in a sequence of topics, via the feature of Modules on Canvas®. Every week a new module was released that included a sequence of 6-8 short videos. This structure was central to the asynchronous online format of the course. Despite the asynchronous online format of the courses, assessment results indicated that students found the collection of short theme-based recordings, coupled with weekly knowledge checks and all other assessments, engaging and efficient.

Coding Assignments Using Jupyter Notebooks and Vocareum. In addition to the recorded lectures, the course material in both courses is heavily based on Python programming, and Jupyter Notebooks are used in the course to present students with an integrated document that contains explanations of concepts along with functioning code. Most lecture notes are provided in a set of Jupyter Notebooks that students can access via GitHub and are made permanently and publicly available so that students can access them even after the course is over.

In addition, both courses use the Vocareum platform for hosting assignments and exams in Jupyter Notebooks. Weekly homework assignments include an auto-graded "skill check" where students complete coding exercises that can be submitted repeatedly for instant feedback, allowing them to keep attempting tasks until they master the skills needed to successfully write and utilize coding tools. There are also homework assignments that include more open-end-

ed questions where students apply their skills to problems similar to what they might encounter in their projects. These homework sets are peer graded, allowing students to see how others solve the problem and get multiple perspectives on how to think about a dataset and optimization formulation, or how to apply different tools. Exams are also based on Jupyter Notebooks hosted on Vocareum and utilize a combination of auto-graded questions and instructor-graded open-ended questions, ensuring that students find the exams consistent in style and content with their prior assignments.

Vertically Integrated Team-Building. The unique environment of classes with mixed internal and external students provides many advantages for students. For example, Georgia Tech students may have broader digital and coding skills, but external students may bring a clearer view of the application opportunities and industrial challenges. A program that allows these groups to interact and learn from each other can synergize their different perspectives and lead to more effective training for all. This environment has allowed us to test the effectiveness of project-based learning in vertically integrated project groups.[24, 25] The groups consist of undergraduate, graduate, and industry students. This structure provides undergraduate students a chance to gain first-hand insight into graduate school and industry careers, and provides industry students with the opportunity to mentor undergraduates and learn from graduate students. The graduate and industry students are responsible for working together to provide the dataset and define the goals for the course project, and all group members are expected to delegate tasks and work together on completing the goals. Importantly, teams are required to meet for at least one hour every week and are encouraged to work together on homework and other assignments, and to complete a bi-weekly group evaluation to ensure that all members are participating and behaving professionally. These requirements help drive engagement by requiring regular, synchronous interactions between students and helps create a cohesive and supportive team where group members can learn from each other.

Project-Based Learning. A key feature of our program is the use of capstone projects in which external students are encouraged to bring real-world data to define group projects. Georgia Tech's experience with industrial collaborations and already in-place master agreements have been leveraged to address potential intellectual property concerns. The outcome is a unique training experience for a new generation of students using real data. The project is completed by "vertically integrated" groups of four or five students (typically one student from industry, two graduate students, and two undergraduate students). The project starts early in the semester when the industry and graduate students are required to provide the dataset for the project and define the goals. Teams are encouraged to work on real datasets and problems from industry or graduate research. (Some examples and outcomes of this approach are provided in the Results and Assessment section of this paper.) This project-driven course design and its focus on real problems help motivate students and provide concrete examples of how the techniques they learn can be applied to real problems. In DACE the project is broken into multiple phases (data preparation, baseline model development, model improvement, and final report/presentation). Similarly, in DDPSE the phases include data preparation or collection, optimization formulation, development of optimization method(s)/algorithm(s), and final report/presentation. In both courses, teams receive feedback from instructors at regular checkpoints. Students can retroactively revise any portion of the project to yield the highest quality results by the end of the semester.

RESULTS AND ASSESSMENT

Example of Project Outcomes When Using Industrial Data Within Vertically Integrated Groups

Project topics range widely but must be related to chemical engineering or chemistry, and often come from real industry or research projects. In DACE some examples include the prediction of polymer solubility, prediction of battery lifetime from early cycle data, analysis of mass spectroscopy data for nanoparticle size detection, prediction of the outputs of unit operations in a cumene production process, and prediction of power generation by a turbine generator. In DDPSE some of the topics studied include optimization of organic electronics, the development of an optimal "Isotherm Modeler" tool, calibration of plasma mass spectrometry data, optimization of hippocampal neurons, optimization of vanadium flow battery design, optimization of phosphoric acid production, design of a biopharmaceutical reactor, building design for minimization of energy consumption, and optimization of air quality in the classroom by scheduling the operation of air purifiers based on real online measurements of air quality.

Ideally, projects provide an opportunity for industry students to bring in a real problem and gain tangible value from the project. A specific example is the prediction of turbine power generation project from Fall 2020, led by participants from The Mosaic Company, which produces phosphate fertilizers. In this project students utilized a variety of regression models to predict the power output of a turbine generator as a function of operating parameters. Industrial students were able to provide real process data, allowing other students in the group to gain experience working with a "data historian" and overcoming challenges associated with cleaning and processing real industry data. The results, shown in Figure 3, showed strong predictive capability and motivated the students to continue working on this project in DDPSE, using the results as the foundation of an internal project at Mosaic that has an estimated gross return of $8 million. In DDPSE, the same students worked on the opti-

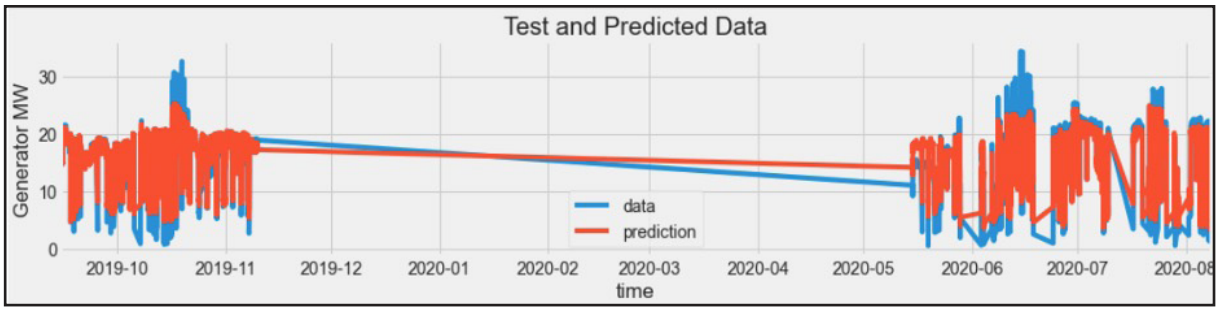

Figure 3. *Example of original data from a turbine power generator at The Mosaic Company (blue), along with the predictions from a data-driven model (red).*

mization of supply chain of phosphate production. In their model they considered costs associated with processing and transportation and used real data to develop correlations that were embedded within the optimization model. In the conclusions of the group's project report assignment they state "*…this study has yielded valuable business insights and a flexible optimization model that can be utilized in the future for deeper investigations into the supply chain of phosphate production…These results can be fed into the larger discussion and investigation of Mosaic's cost model and provide the ability for the mine planning department to make data driven decisions*."

This project highlights how the project-driven curriculum connects the two core courses of the data science certificate program and drives engagement with students and industry. The highlighted example project also shows that there is continuity and cohesiveness of topics built in the two core courses. This allows students to learn about how to process and analyze their data as well as how to use the techniques from DACE to make optimal decisions or solve inverse problems. In other words, the two core courses are designed to "close the loop" between data collection, processing, analysis, and decision-making.

This project-driven format has also led to success stories on the academic front. Specifically, one of the graduate students provided data from his research, which led to a successful peer-reviewed publication.[26] In this publication, literature data on processing and mobility properties of organic field effect transistors were used to develop a classification model to aid in identifying optimal design regions of polymer concentrations. Results of this data-driven model were confirmed by experiments.

Visualization and Analysis of High-Dimensional Data Using the Dow Dataset

One strategy used to make the course more accessible and engaging for chemists and chemical engineers is the use of datasets relevant to chemical engineering and related discipline. This diverges from typical computer science courses and boot camps, where examples often focus on generic toy datasets, text processing, or image analysis. In the DSCI courses, we supplement standard datasets such as MNIST [27] and UCI ML Repository,[28] with examples including analysis of infrared spectra, prediction of materials properties, and correlating process inputs and outputs for chemical processes. One specific example, generously provided by Dow Chemical, is a dataset of approximately 10k points in a time series that correlates approximately 40 operating parameters to the impurity level from a series of distillation columns (Figure 4a).

The Dow dataset is provided as a raw spreadsheet that includes common artifacts and missing values, providing an opportunity for students to learn to "wrangle" data into well-organized structures suitable for analysis. Students also use supervised regression techniques to make quantitative prediction of process outputs as a function of process inputs and learn to assess the quality and accuracy of the resulting models (Figure 4b). Moreover, students use techniques such as dimensional reduction and clustering to analyze the data and to identify different operating regimes, including outlier detection (Figure 4c). These experiences help students connect the various data science techniques back to concepts from chemical engineering such as start-up and shutdown, flow rates, and impurity levels.

Assessment of Learning Outcomes and Student Experience

When creating this new program, we defined two objectives in assessing student success:

- Objective 1: Competence in Fundamentals of Data Science in Chemical Manufacturing. Students will develop and demonstrate competence in fundamental knowledge of data science relevant to chemical manufacturing.
- Objective 2: Competence in Software Applications of Data Science in Chemical Manufacturing. Students will develop and demonstrate competence in using programming and software that uses data science to solve real-world problems in chemical manufacturing.

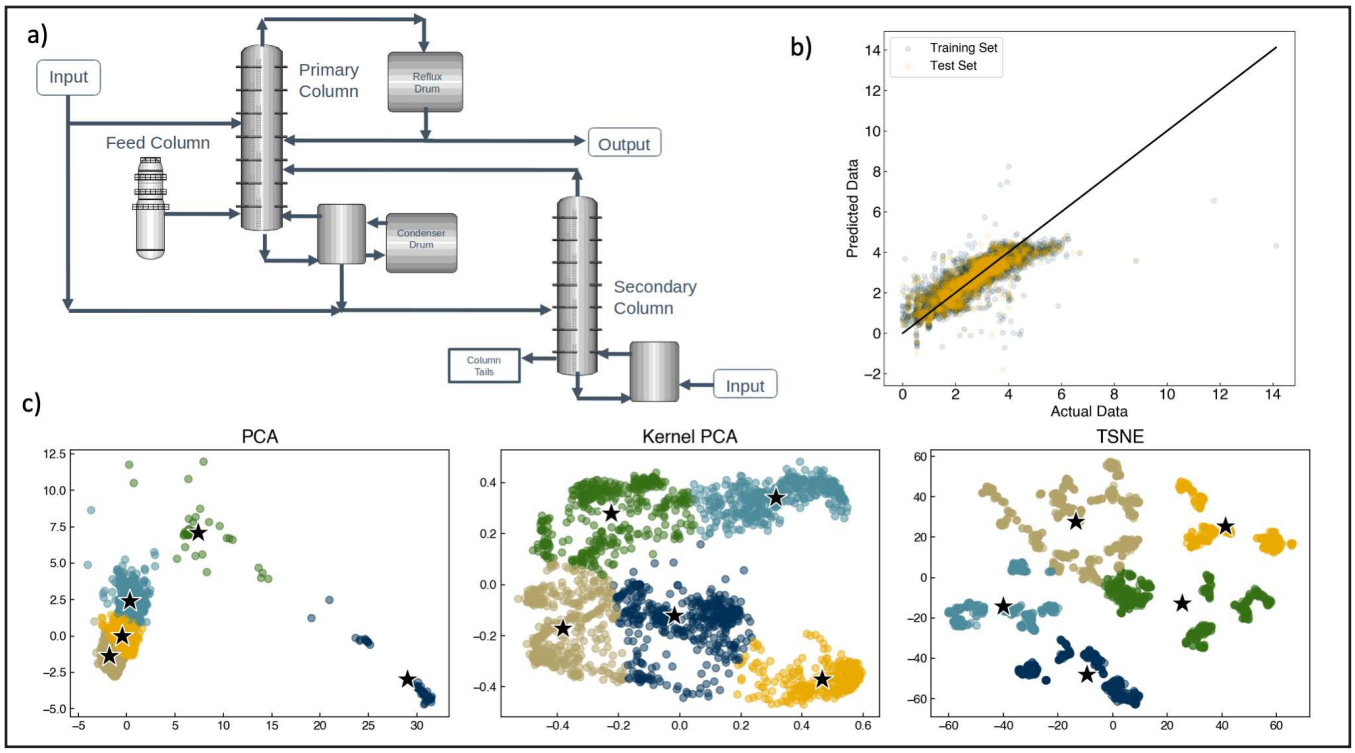

Figure 4. *Examples of (a) a process flow diagram for the Dow dataset, (b) a parity plot visualizing the results of a regression model, and (c) a comparison of multiple types of dimensional reduction and clustering analyses, where different colors represent different "clusters" of operating conditions, and the stars represent the centroid of each cluster.*

In the first year of the program, to evaluate Objective 1, the score on the final exam of DACE was analyzed to determine the number of students earning 80% or higher, with a target of 75% of students meeting this criterion. For the internal Georgia Tech students, 88% (n above 80% = 15, total n = 17) of the learners received a final exam score of 80% or higher in DACE. For the external online students, 86% (n above 80% = 6, total n = 7) of the learners received a final exam score of 80% or higher in DACE.

For Objective 2, the score on the final exam of DDPSE was similarly analyzed to determine the number of students earning 80% or higher, again with a target of 75% of students meeting this criterion. Among internal Georgia Tech students, 76% (n above 80% = 16, total n = 21) of the learners received a final exam score of 80% or higher in DDPSE. For the online students, 86% (n above 80% = 6, total n = 7) received a final exam score of 80% or higher in DDPSE. Thus, in the first year of the program, both the internal and the external cohorts performed above the target level in both objectives.

Mid-semester surveys were also conducted for external students in the first cohort of the program. Six of the seven students responded to the survey. When asked "What was your primary motivation for obtaining this graduate certificate?," four students indicated that it was to expand their knowledge base. One selected "change jobs/get a new job," and one selected "other." When asked, "What is your current employment status?," five students indicated that they were working full-time while one was a student.

Students were also asked, "So far, how would you rate your satisfaction with the following aspects of course delivery?" The results are shown in Figure 5. Overall, students were most satisfied with the online lectures, the quality of feedback from instructors, and the number of opportunities to interact with course instructors. Since these are online courses, it is particularly notable that the students are satisfied with their interactions with the instructors. The students reported being less satisfied with the number of opportunities to interact with fellow students, and that concern is being considered in current offerings of the course.

Another question dealt with instructional technology, as shown in Figure 6. Students were most satisfied with Canvas and least satisfied with Honorlock™, the software used for virtual proctoring. In subsequent semesters the instructors reduced or even eliminated the use of proctoring software and incorporated honor code statements into exams.

Students were also asked, "So far, how would you rate your experience in the Online Graduate Certificate in Data Science for the Chemical Industry?" Three students indicated that the program exceeded or far exceeded their expecta-

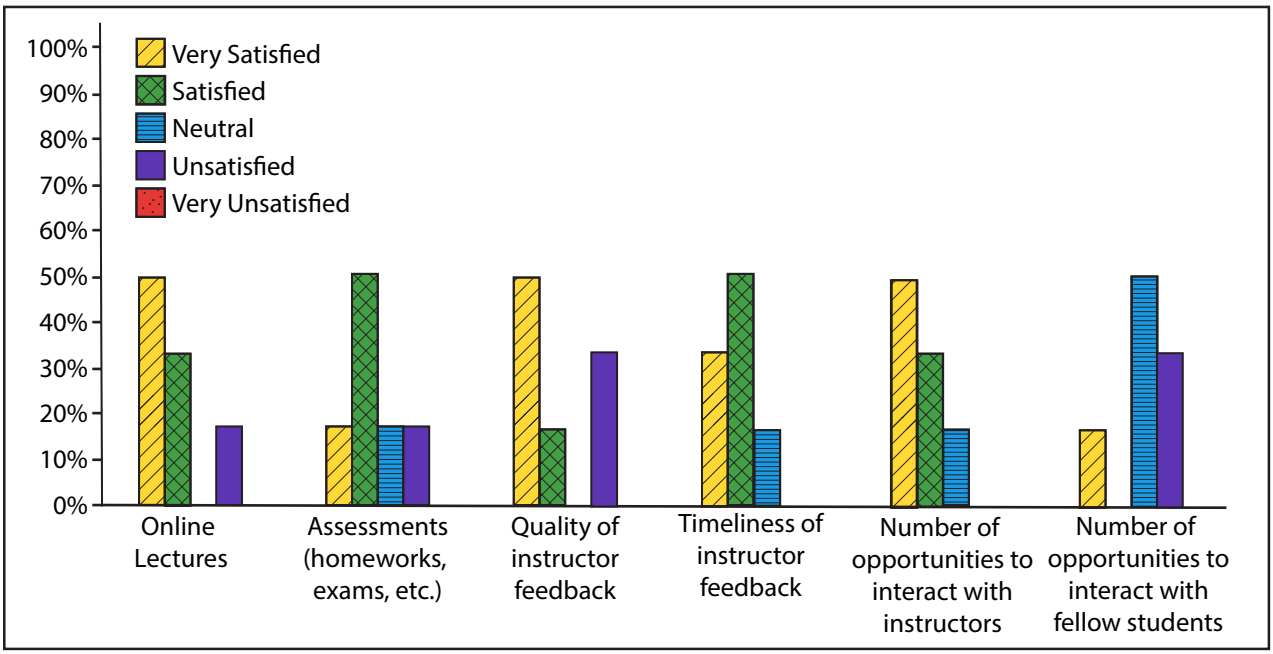

Figure 5. *Response to "So far, how would you rate your satisfaction with the following aspects of course delivery?"*

tions, two students indicated that it met expectations, and one student indicated that it fell short of expectations. Based on this survey together with direct student feedback, the program is continuing to build upon its successful early performance to serve student needs.

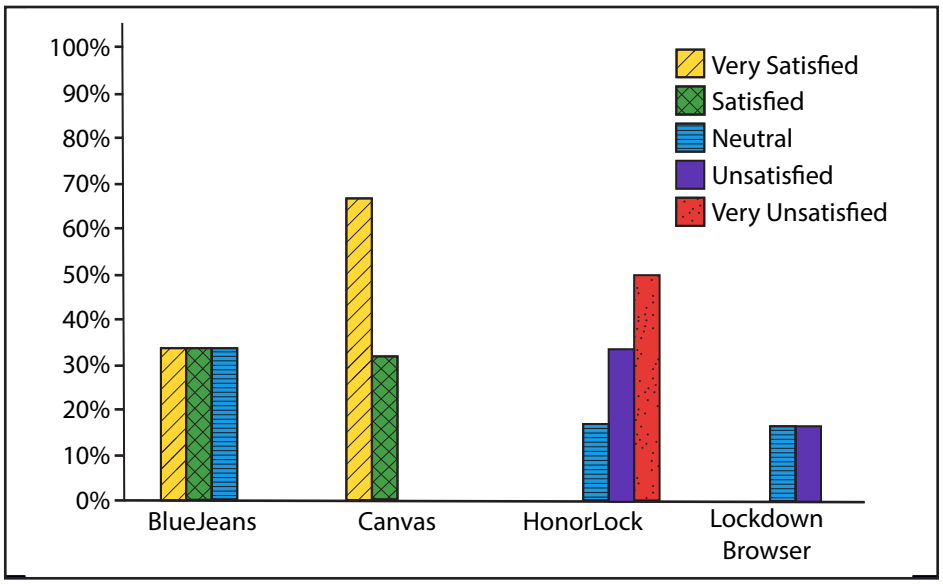

Figure 6*. Responses to "Please rate your satisfaction with the following course technologies."*

DISCUSSION

As one of the first stand-alone graduate certificates at Georgia Tech, this program can be a model for others to adopt, partnering with professional education for this online asynchronous format. The Graduate Certificate in Data Science for the Chemical Industry is Georgia Tech's first fully online graduate certificate. By serving both off-campus and Georgia Tech students, it creates a template for much broader innovation in graduate education that is likely to appeal to a large group of students for whom a complete online MS is too large of a commitment. The program also has the potential for improving diversity, equity, and inclusion in chemical engineering. The School of Chemical & Biomolecular Engineering at Georgia Tech graduates one of the largest and most diverse student populations in the discipline in the US,[29] and the asynchronous online format allows for improved inclusion of non-traditional and continuing education students who may be juggling education with other family or work responsibilities. Although the current demographics of the chemical industry do not reflect the diversity of society, a testimonial from at least one participant who was a member of an under-represented group from industry indicated that the knowledge they gained within the first year of the program enabled them to pursue leadership roles

in new data-science related careers within the company. This example suggests that the program may help advance the careers of existing under-represented industry professionals. Nonetheless, we recognize that there is considerable room for improvement in encouraging participation of under-represented students, and we strongly encourage participation from under-represented groups both at Georgia Tech and within the chemical industry.

The diversity of the student backgrounds and perspectives is a strength of the program, especially for the vertically integrated group projects. However, it also means that some students may struggle, for example, with programming in the first course. Notably, the diversity in skill level is relatively consistent across the various types of students. Some undergraduates are double-majoring with computer science, while others have only had basic training in programming. Similarly, some industry professionals are actively working on data science projects, while others have not had a programming course in many years. While this may shift as undergraduate curricula evolve, we anticipate significant diversity in skill level across all student groups to be a challenge for the foreseeable future. The asynchronous format of the course provides unique possibilities for individually tailoring the curriculum to each student. As a first step, we recommend that students take an optional Python "boot camp" before beginning the first course if they do not already have experience with Python. In addition, we track student engagement throughout the semester using a combination of multiple choice "knowledge checks" and auto-graded programming "skill checks" that allow students to submit their code an unlimited number of times until it is correct. These checks often drive engagement at weekly office hours, and students who struggle are encouraged to set up individual meetings with instructors or TAs as needed. In addition, the mandatory weekly group meetings help ensure that students remain engaged with the course and provide them with peer support if they are struggling with coding basics. The relatively high retention rate between the two courses (Figure 1) indicates that these support strategies are successful.

We feel that the current DSCI program serves the needs of both chemical industry professionals and Georgia Tech students well and helps fill an educational gap in the current standard curriculum of chemical engineering education. We note that the DSCI courses generally focus on more "basic" concepts and forego recent and emerging advances such as deep learning. This was an intentional decision, with the goal of building a strong conceptual framework and providing a broad overview, rather than attempting to train students in the most advanced techniques. However, we also recognize that the field is still evolving. One potential disadvantage of the online asynchronous format is the relatively static nature of pre-recorded lectures. This can be mitigated in the short term by revising individual topics/lectures as needed or adding additional supplemental modules that can address urgent educational needs identified by industry participants. For example, towards the end of the DDPSE course, advanced topics are covered, such as physics-informed ML and how that is linked to the optimization theory taught earlier in the course, as well as the concept of deep neural networks and why/when one would use them. This material did not exist in the first offering of the course and was added in the second year based on new research and feedback from students. In the longer term, the development of new courses, or the full re-design of the existing course curriculum may be necessary to address the changing skillsets of incoming students or to teach new skills that arise as machine learning and artificial intelligence are more deeply embedded into the chemical process industry. At the same time, we also note that the programming and data-analytics skills provided to the students can serve as an entry point to more advanced training or alternative careers outside of the chemical process industry. We have seen examples of such students who graduated from the program and continued to MS degrees in computer science, secured jobs in consulting, and even worked for start-up companies for block-chain technology. This indicates that the skills that chemical engineering students learn through these courses are flexible and will likely enable graduates from the DSCI program to adapt to the changing needs of the field without the need for additional (re)training.

CONCLUDING REMARKS

Georgia Tech's DSCI provides a pathway for industry professionals to learn data science in the context of the chemical industry, in an asynchronous online format that is conducive for working professionals. The vertically integrated project groups allow internal and external students to collaborate on projects involving real data, bringing together their diverse perspectives and skills.

ACKNOWLEDGMENTS

The authors thank Chris Jacobs (3M), Leo Chiang (Dow), and Zak Kuiper (Mosaic) for their formative input and contribution of data sets. The authors would also like to thank the Georgia Tech Professional Education group for their hard work creating high-quality video lectures, their continuous support of teaching tools via Canvas, and their expertise on effective online course design. Finally, the authors would like to thank the Georgia Tech and the ChBE administration (David Sholl, Carson Meredith, and Christopher Jones) for their continued support.

REFERENCES

1. Fernández L. U.S. Chemical industry - statistics & facts. https://www.statista.com/topics/1526/chemical-industry-in-the-us/. Accessed August 19, 2022.
2. Chui M, Manyika J, Miremadi M, Henke N, Chung R, Nel P, and Malhotra S (2018) *Notes from the AI frontier: Applications and value of deep learning*. https://www.mckinsey.com/featured-insights/artificial-intelligence/notes-from-the-ai-frontier-applications-and-value-of-deep-learning Accessed August 25, 2022.
3. Chiang L, Lu B, and Castillo I (2017) Big data analytics in chemical engineering. *Annual Review of Chemical and Biomolecular Engineering*. 8(1):63-85. DOI: 10.1146/annurev-chembioeng-060816-101555.
4. Duever TA (2019) Data science in the chemical engineering curriculum. *Processes*. 7(11):830.
5. Feise HJ and Schaer E (2021) Mastering digitized chemical engineering. *Education for Chemical Engineers*. 34:78-86. DOI: https://doi.org/10.1016/j.ece.2020.11.011.
6. The National Academies of Sciences, Engineering and Medicine (2018) *Data Science for Undergraduates*.
7. The National Academies of Sciences, Engineering, and Medicine (2022) *New Directions for Chemical Engineering*.
8. Venkatasubramanian V (2019) The promise of artificial intelligence in chemical engineering: Is it here, finally? *AIChE Journal*. 65(2):466-478. DOI: https://doi.org/10.1002/aic.16489.
9. Georgia Tech - Graduate Certificate in Data Science for the Chemical Industry. https://www.chbe.gatech.edu/data-science-certificate. Accessed August 19, 2022.
10. University of Washington - Advanced Data Science Option. https://www.cheme.washington.edu/graduate_students/prosp_grad/program/ADS.html. Accessed August 19, 2022.
11. Columbia University - MS in Chemical Engineering with Concentration in Data and Computational Science. https://www.cheme.columbia.edu/ms-chemical-engineering-concentration-data-and-computational-science. Accessed August 19, 2022.
12. Purdue University - Data Science MS. https://engineering.purdue.edu/ChE/academics/graduate/masters/datascience-concentration. Accessed August 19, 2022.
13. Cornell University - Computational Informatics. https://www.cheme.cornell.edu/cbe/academics/graduate-programs/meng/specializations/computational-informatics. Accessed August 19, 2022.
14. University of Michigan - Data Science Certificate. https://midas.umich.edu/certificate/. Accessed August 19, 2022.
15. Lee JH, Shin J, and Realff MJ (2018) Machine learning: Overview of the recent progresses and implications for the process systems engineering field. *Computers & Chemical Engineering*. 114:111-121. DOI: https://doi.org/10.1016/j.compchemeng.2017.10.008.
16. Medford AJ, Kunz MR, Ewing SM, Borders T, and Fushimi R (2018) Extracting Knowledge from Data through Catalysis Informatics. *ACS Catalysis*. 8(8):7403-7429. DOI: 10.1021/acscatal.8b01708.
17. Pistikopoulos EN, Barbosa-Povoa A, Lee JH, Misener R, Mitsos A, Reklaitis GV, Venkatasubramanian V, You F, and Gani R (2021) Process systems engineering–the generation next? *Computers & Chemical Engineering*. 147:107252.
18. Qin SJ (2014) Process data analytics in the era of big data. *AIChE Journal*. 60(9):3092-3100. DOI: https://doi.org/10.1002/aic.14523.
19. Schmidt J, Marques MRG, Botti S, and Marques MAL (2019) Recent advances and applications of machine learning in solid-state materials science. *npj Computational Materials*. 5(1):83. DOI: 10.1038/s41524-019-0221-0.
20. Udugama IA, Gargalo CL, Yamashita Y, Taube MA, Palazoglu A, Young BR, Gernaey KV, Kulahci M, and Bayer C (2020) The role of big data in industrial (bio)chemical process operations. *Industrial & Engineering Chemistry Research*. 59(34):15283-15297. DOI: 10.1021/acs.iecr.0c01872.
21. Westmoreland PR (2014) Opportunities and challenges for a Golden Age of chemical engineering. *Frontiers of Chemical Science and Engineering*. 8(1):1-7. DOI: 10.1007/s11705-014-1416-z.
22. Hart WE, Laird CD, Watson J-P, Woodruff DL, Hackebeil GA, Nicholson BL, and Siirola JD (2017) *Pyomo-Optimization Modeling in Python*. Springer. New York, NY.
23. Verrett J, Boukouvala F, Dowling A, Ulissi Z, and Zavala V (2020) Computational notebooks in chemical engineering curricula. *Chemical Engineering Education*. 54(3):143-150. https://journals.flvc.org/cee/article/view/116661 Accessed August 19, 2022.
24. Baxter M, Byun B, Coyle EJ, Dang T, Dwyer T, Kim I, Lee C, Llewallyn R, and Sephus N (2011) On project-based learning through the vertically-integrated projects program. *Proceedings 2011 Frontiers in Education Conference (FIE)*. DOI: 10.1109/FIE.2011.6143064.
25. ElZomor M, Mann C, Doten-Snitker K, Parrish K, and Chester M (2018) Leveraging vertically integrated courses and problem-based learning to improve students' performance and skills. *Journal of Professional Issues in Engineering Education and Practice*. 144(4):04018009. DOI: doi:10.1061/(ASCE)EI.1943-5541.0000379.
26. Venkatesh R, Zheng Y, Viersen C, Liu A, Silva C, Grover M, and Reichmanis E (2021) Data science guided experiments identify conjugated polymer solution concentration as a key parameter in device performance. *ACS Materials Letters*. 3(9):1321-1327. DOI: 10.1021/acsmaterialslett.1c00320.
27. Deng L (2012) The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Processing Magazine*. 29(6):141-142.
28. Dua D and Graff C. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml. Accessed August 19, 2022.
29. ASEE. American Society for Engineering Education. https://ira.asee.org/profiles-of-engineering-engineering-technology/#:~:text=ASEE%20publishes%20the%20leading%20data,%26%20Engineering%20Technology. Accessed August 19, 2022. ❐