MRS Advances © 2020 Materials Research Society
DOI: 10.1557/adv.2020.171

Merging Materials and Data Science: Opportunities, Challenges, and Education in Materials Informatics

Thomas J. Oweida[1], Akhlak Mahmood[1], Matthew D. Manning[1], Sergei Rigin[1], Yaroslava G. Yingling[1]

[1]*Department of Materials Science and Engineering, North Carolina State University, 911 Partners Way, Raleigh, NC 27695-7907, USA*

ABSTRACT

Since the launch of the Materials Genome Initiative (MGI) the field of materials informatics (MI) emerged to remove the bottlenecks limiting the pathway towards rapid materials discovery. Although the machine learning (ML) and optimization techniques underlying MI were developed well over a decade ago, programs such as the MGI encouraged researchers to make the technical advancements that make these tools suitable for the unique challenges in materials science and engineering. Overall, MI has seen a remarkable rate in adoption over the past decade. However, for the continued growth of MI, the educational challenges associated with applying data science techniques to analyse materials science and engineering problems must be addressed. In this paper, we will discuss the growing use of materials informatics in academia and industry, highlight the need for educational advances in materials informatics, and discuss the implementation of a materials informatics course into the curriculum to jump-start interested students with the skills required to succeed in materials informatics projects.

Downloaded from https://www.cambridge.org/core. NC State University Libraries, on 24 Jul 2020 at 17:42:12, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1557/adv.2020.171

Downloaded from https://www.cambridge.org/core. NC State University Libraries, on 24 Jul 2020 at 17:42:12, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1557/adv.2020.171

MATERIALS INFORMATICS IN ACADEMIA

Materials-based research has adopted the use of machine learning (ML) as an analytical tool. ML encompasses any algorithm whose performance will improve, or learn, as it is exposed to or trained on larger quantities of quality data. Implementing these ML algorithms with a specific workflow to overcome the unique challenges of materials-based research is termed materials informatics (MI). The application of ML tools to materials science data and the use of MI workflow to design new materials and techniques has shown exponential growth with over 2,000 publications during the past decade (Figure 1). The United States leads the global effort in MI with almost half of these publications. To date, most of the publications have largely focused on facilitating materials design, parameterizing potentials for in silico techniques, and optimizing materials characterization techniques [1]–[6]. For example, researchers have started to employ ML algorithms to process undetected or complex trends in databases containing first principle calculations data [7]. Ultimately, this has led to the proposal and synthesis of promising surface coatings [8], alloys[9]–[12], perovskites [13], and composites [2] that meet specified target properties for a specific application. MI has not only been useful in designing and predicting properties of new materials, but has also been vital in the recent development of new potentials used for in silico approaches via rapid parameterization [14], [15] at a reduced computational expense [15], [16] and the development of completely data driven potentials [17], [18]. These advancements have been utilized to push the length and time scales of the

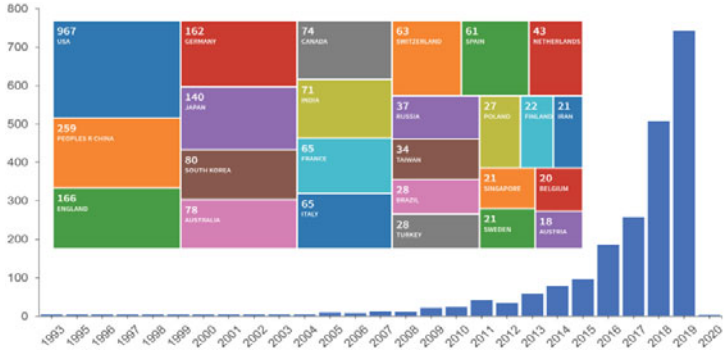

Figure 1. Illustrates the recent boom in publications per year that combine aspects of data science into materials-based research. The inset shows the breakdown of materials informatics publications by country. Results are based off 2,268 records for TOPIC: ("materials informatics") OR TOPIC: ("machine learning" and "materials"). <<color online>>

current capabilities of simulations while maintaining the same level of accuracy as higher resolution simulation techniques. The MI framework to develop these new potentials has already been laid out in multiple studies [19]–[22]. For example, Huan et al. discuss a universal approach for creating atomistic force fields via ML [22]. MI has also been used to analyze phenomenological parameters such as the work performed by Miles, Leon, Smith, and Oates that looked at the uncertainty and sensitivity of parameters in a ferroelectric continuum model for lead titanate [23], [24]. Experimental characterization techniques are also a beneficiary of MI approaches. For example, a Bayesian inference approach was shown to provide many advantages for X-ray diffraction peak fitting over traditional approaches such as Reitveld refinement. Specifically, the Bayesian approach has the ability to escape from false minima, incorporate prior knowledge of the material into analysis, and provide uncertainty quantification [25]. MI has also been used to analyze position averaged convergent beam electron diffraction patterns with a convolutional neural network that achieved great speed and accuracy compared to brute force methods [26]. Overall, the rate of adoption of MI workflow to speed up characterization, simulations and materials discovery has been remarkable.

MATERIALS INFORMATICS IN INDUSTRY

The use and application of ML tools has rapidly grown in industry with CrunchBase listing over 5,000 start-ups that implement data science tools to carry out the production of their products and services. This has resulted in an explosion of spending on ML and artificial intelligence (AI) with the International Data Corporation predicting an investment of $57.6B by 2021 [27]. Among these start-ups are Materials Informatics companies like Citrine Informatics which has more than doubled in size over the last 2 years [28]. Citrine Informatics aims to accelerate materials discovery through their proprietary ML platforms that stores and uses materials data from its partnerships with universities, national labs, and corporations. With the rise of companies like Citrine Informatics, engineers and scientists that possess both knowledge of domain science (like materials science and engineering) and data science techniques are in very high demand.

The incentives to pursue educational and research opportunities in MI remains high as the fields of MI and artificial intelligence continue to see significantly above-average job growth. In 2019, job postings on LinkedIn for data scientists and machine learning engineers increased by 56% and

Downloaded from https://www.cambridge.org/core. NC State University Libraries, on 24 Jul 2020 at 17:42:12, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1557/adv.2020.171

96% respectively [29]. The Bureau of Labor Statistics predicts that the number of jobs in Mathematics and Statistics (which includes data scientists) will rise by 30% from 2018 to 2028 [30]. In addition to job growth, informatics jobs provide above average compensation. LinkedIn users reported median salaries of $130,000 and $182,000 for the data scientist and machine learning engineer positions and data scientists who handle unstructured data and have coding skills earn up to 34% more than other analysts. This group is also nearly three times as likely to have a degree in engineering or science (34% vs 13%) [31]. As the use of data science tools matures and spreads to different industries and domains, the demand for professionals with domain knowledge and the ability to handle incomplete and heterogeneous data is increasing. This trend is especially applicable to materials science, where structured databases, off-the-shelf data analytics software packages and professionals with both domain and data science knowledge are in short supply.

CHALLENGES

MI is undeniably a valuable tool for materials scientists as the modern pace of materials innovation has become intractable by traditional approaches. The success of MI in academia and industry has only reinforced this truth as structure and property predictions across vast chemical spaces become simultaneously cheaper and more accurate. However, materials informatics is still lagging behind other fields that have adopted data science approaches due to the unique challenges inherent to materials datasets. One of the most impactful processes in each MI approach is the user-dependent choice of material descriptors. In general, these descriptors need to sufficiently identify unique atomic environments, while being invariant to transformations such as translation, rotation, and permutations of like elements [17]. However, these descriptors can quickly become computationally expensive, which is especially true for soft matter as the exploration space is inherently highly dimensional [32], [33]. These materials can have properties heavily reliant upon this design space as their sequence, environment, length, chemical composition, density, etc. can drastically change morphology and non-bonded interactions [34]–[36]. Thus, developing a framework that can identify the optimal material descriptors for each MI application can help overcome one of the biggest barriers that has kept MI from realizing its full potential. Early works have already targeted this issue through the development of standard notation such as SMILES for molecules and

Downloaded from https://www.cambridge.org/core. NC State University Libraries, on 24 Jul 2020 at 17:42:12, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1557/adv.2020.171

BigSMILES for macromolecules [37]–[39]. In addition, some works have performed analysis on the influence of numerous materials descriptors ranging from crystal chemistry to electronic structure descriptors used to predict multiple properties of intermetallic compounds [11].

Databases would seem to be the easy solution to standardize the structure of reported data, however, current databases are for particular purposes or limited to specific materials class(es) limiting their viability for use in MI studies. In addition, most databases do not report material processing details resulting in a possible disconnect between the structure-property relationship that is fundamental to material science. This is especially important for less-ordered materials commonly found in soft materials and glasses. For more information on the additional challenges for disordered materials we recommend referencing "Soft Matter Informatics: Current Progress and Challenges" by Peerless et al. [33] and "Data-driven glass/ceramic science research: Insights from the glass and ceramic and data science/informatics communities" by De Guire et al. [40].

Collecting data from previous publications also possess significant challenges. In a recent paper, the quality of data reported was highly concerning for inorganic materials synthesis recipes. Through a text mining approach, it was found that the overall extraction yield was 28% of total papers. Out of the successfully mined publications, 30% of papers did not contain a complete set of starting materials and final products, thus reconstruction of the reaction was not possible. Lastly, 42% of potential reactions were not reconstructed due to an incomplete or overcomplete set of extracted precursor/target materials, or a failure to correctly parse chemical composition [41]. Thus, the already limited materials datasets are further reduced in size due to poor data quality and lack of standards for reporting data.

For future database development, the materials informatics community needs to follow the examples of well-established databases such as the Protein Data Bank (PDB) [42]–[44]. The development of the PDB has created a culture, incentives and level of prestige that benefits each researcher that successfully submits a protein structure in this database for others to use. This effective data sharing in PDB database resulted in the growth and development of structural bioinformatics field. In addition to centralizing large amounts of data, PDB has implemented a data quality metric that ensures only quality data exists within the database effectively

Downloaded from https://www.cambridge.org/core. NC State University Libraries, on 24 Jul 2020 at 17:42:12, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1557/adv.2020.171

reducing the burden and time it takes to pre-process and filter the data for analysis.

For databases specific to materials science, The National Institute of Materials Science (NIMS) deserves special recognition as being one of the front runners for database development for MI applications [45]. NIMS is the co-copyright owner of databases such as the Pauling Files [46] which provides reliable data on crystal structure and phase diagrams in addition to being the owner of other respected databases such as PoLyInfo which provides data for polymeric materials design [47]. These databases are traditionally curated by hand as NIMS employees comb through literature daily assessing accurate information for entry into a database. Thus, NIMS curation of data has resulted in the development of databases that have been successfully used as the source of information in numerous MI studies.

EDUCATION

In a recent report from TMS, it was reported that only 9 out of the 50 sampled Materials Science programs offered a course that referenced "data science," "data handling," or the utilization of "databases" [48]. This creates a disconnect between the skills of graduating students and the desires of employers who seek more interdisciplinary training among materials graduates [49]. One of the root causes for the lack of education in materials informatics may be the shortage of faculty suited to teach the course. The limited resources, textbooks and course models currently available to faculty serves as a barrier for the induction and development of new MI classes. In the remainder of this section, we will discuss the course format and education opportunities successfully launched at the Materials Science and Engineering Department at North Carolina State University. While this text will serve as a resource and discuss a current course model, there is a need to develop a viable course textbook as we find current textbooks to either focus too much on the computer science aspects of ML or fail to address the unique challenges associated with MI as discussed in the previous section.

In order to develop the curriculum that prepares students for MI field, it is critical to understand the basic skillset a student needs to comfortably understand and incorporate materials informatics into future work, whether in academia or industry. Figure 2 highlights these essential skills which include math and statistics to understand protocol for handling different types and sizes of data, databases to help store and collect materials

Downloaded from https://www.cambridge.org/core. NC State University Libraries, on 24 Jul 2020 at 17:42:12, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1557/adv.2020.171

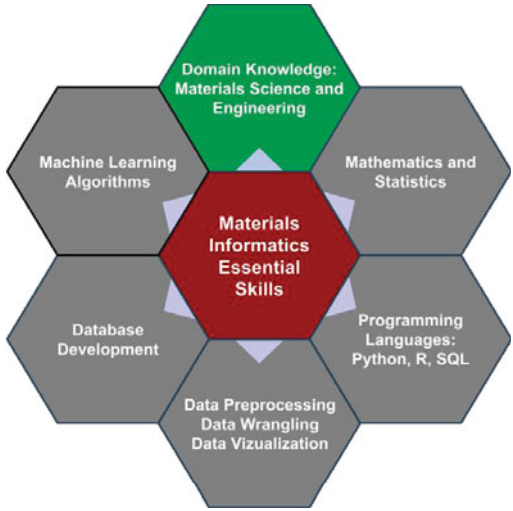

Figure 2. Highlights the necessary skillset for materials informatics. The green hexagon represents the current information that is taught in materials science programs and the grey hexagons represent required knowledge outside of materials science and engineering. <<color online>>

data, data wrangling and pre-processing to prepare data for analysis, coding/machine learning algorithms to understand and run the MI process, and most importantly materials science domain knowledge. This is not to say materials graduates must be experts in all these fields, but they must have enough of an interdisciplinary background to understand current literature and interface with experts in informatics to exploit their own materials expertise.

At North Carolina State University, there has been a large effort to enable students with the proper training in materials informatics through the NSF funded Data-Enabled Science and Engineering of Atomic Structure (SEAS) program [50]. Specifically, the SEAS program has created incentives for students through fellowships, seminars, group discussions, hackathons, and educational opportunities through the establishment of Graduate Certificate in Materials Informatics (GCMI), and networking opportunities that connect students with universities, national labs, and companies interested in hiring materials students with knowledge of informatics. The GCMI is designed for interdisciplinary graduate education at the intersection of materials science, engineering, and data science with the aim of preparing the next generation of materials engineers given the growing demand for data-science skills and knowledge of artificial

Downloaded from https://www.cambridge.org/core. NC State University Libraries, on 24 Jul 2020 at 17:42:12, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1557/adv.2020.171

intelligence. The skills and knowledge obtained here will serve as foundation for the understanding of materials informatics and high throughput materials discovery that will improve a student's career prospects. The core course for CGMI is Materials Informatics which was designed to enable students with the practical implementation of machine learning techniques to various materials science problems and to introduce the MI required skills in Figure 2. Namely, the course covers an overview of materials knowledge, data management, and machine learning, while in-depth statistics and coding are referenced in concepts throughout the course but taught in other courses within the curriculum. While the course includes some basic review of necessary math and statistics, we do recommend that the students enrolled in this course already be familiar with main concepts in statistics and uncertainty quantification. The hands-on implementation of the course is based on (1) reproducing the recently published data and (2) application of learned techniques to the student-driven project. For the known example, we highly recommend the paper and supplementary information in the publication "Predicting the thermodynamic stability of perovskite oxides using machine learning models" by Li, Jacobs, and Morgan [51]. This paper shares the code and data used to form the reported results which will allow students and faculty to recreate the study and verify their correct implementation of that specific MI approach by comparing their results with those reported in the literature. In addition, the 'scikit-learn' python package utilized by Li, Jacobs, and Morgan provides an excellent description of data science tools with examples. The complete course flow is illustrated in Figure 3a.

Materials and Data Collection (Weeks 1 and 2)

The course begins with the basic review of statistics and python basics followed by an overview of the typical data obtained from materials characterization (Fig. 3a). This encourages students to think about the raw data structure obtained through various materials characterization techniques across subdisciplines of materials science. For example, the differences in data structure and resolution can be discussed for characterization techniques such as TEM, X-ray diffraction, and computational techniques. During this portion of the class, it is important to note the common occurrence of small datasets and high dimensional design spaces in materials science. It should be stressed that the understanding of the particular material structure and properties is based on the physics infused into the dataset, limiting the benefits of MI approaches that do not maintain this information. Outside of

Downloaded from https://www.cambridge.org/core. NC State University Libraries, on 24 Jul 2020 at 17:42:12, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1557/adv.2020.171

class, students should be thinking about a materials project that they would like to work on throughout the semester. Figure 3(b) shows an example of a student's project idea to predict self-assembled morphologies based on the design of amphiphilic diblock copolymers. During this period, Homework 1 is assigned where students should think of all possible characterization techniques that can be used to investigate the properties of selected material in their project idea. The students will also choose one characterization techniques and list all the tunable material preparation and characterization parameters that can influence results in their proposed project. Figure 3(b) shows an example of a flow diagram from a student's homework. The student listed 8 techniques to study micelles in solution and settled on describing the parameters of a computational method in detail.

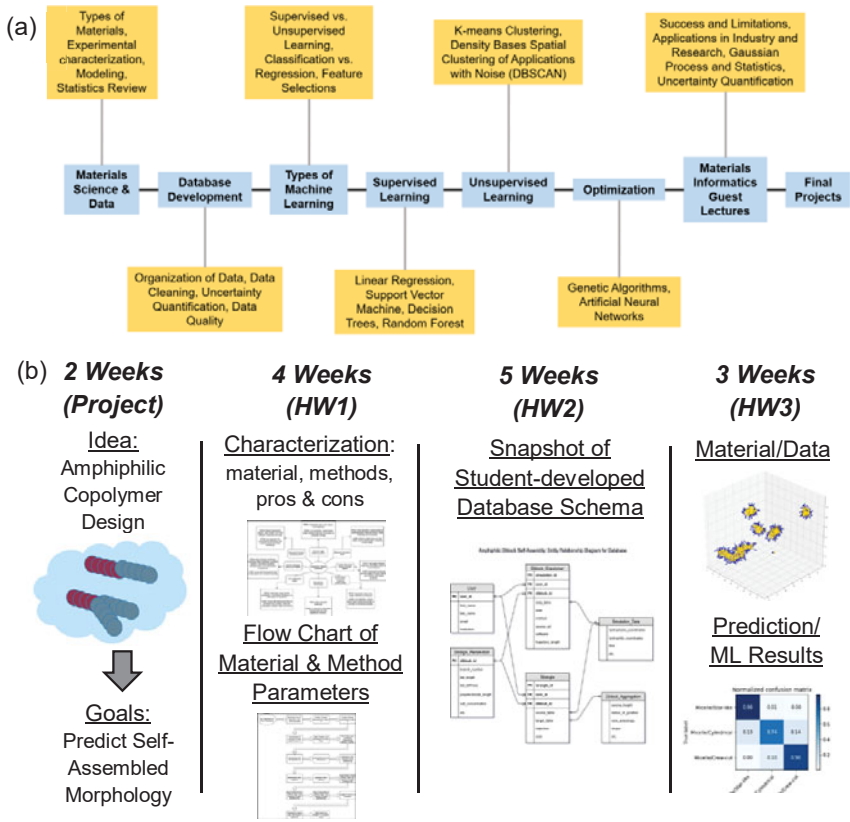

Figure 3. (a) Represents the syllabus topics of the course (blue) and suggested subtopics (yellow). (b) Represents the timeline for each out of class assignment along with an example of a student's homework assignment figures as discussed in the text. <<color online>>

Downloaded from https://www.cambridge.org/core. NC State University Libraries, on 24 Jul 2020 at 17:42:12, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1557/adv.2020.171

Downloaded from https://www.cambridge.org/core. NC State University Libraries, on 24 Jul 2020 at 17:42:12, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1557/adv.2020.171

Data wrangling and database design (Weeks 3-6)

The course then transitions into discussing database structures which teaches students about the organization of data, data processing, data quality, and the value of uncertainty quantification. This data management step prepares students to pre-process data that can be fed into ML algorithms for materials prediction and optimization problems. This section of the course directly builds upon the understanding of data for various materials science and engineering problems, which was covered in weeks 1 and 2. During this time, students are assigned Homework 2, which requires data collection, wrangling and database design for both the example data from Li et al [51]. and student project. Ultimately, this homework culminates in the sketching of a database that represents the workflow of the important material and characterization parameters in Homework 1. Figure 3(b) shows a model database developed by a student. The database information should sufficiently maintain the structure-process-property relationship for each material entered.

Machine learning techniques (Weeks 7-11)

Once students are comfortable with data collection, data wrangling, and data organization concepts, an overview on the types of ML algorithms is discussed in class. This includes teaching students the difference between supervised and unsupervised learning, the difference between classification and regression problems, and highlights the importance of feature selection. The following weeks of lecture then discuss a few algorithms in detail, teaching students the basics of how they work and providing examples through in-class, hands on examples. The in-class examples were based off of the 'scikit-learn' python package but could be altered to the instructors choosing. Figure 3(a) provides some suggested algorithms to explain in detail based off of their current usage in MI research. Once the students begin learning about specific algorithms, Homework 3 is assigned to reproduce and work with data in example paper Li et al. [51] and to utilize a ML algorithm of their choice on the dataset they have chosen for their final project.

Implementation of MI (Weeks 12-15)

The final 3 weeks of class consist of lectures on state-of-the art implementation of MI to different problems, discussion on bottlenecks in various fields of materials, and guest speakers from academia or industry.

The lectures focus on some of the challenges discussed in the 'challenges' section and provide guidance on how the MI community can overcome some of these challenges. The guest speakers discuss their current usage of materials informatics providing students with an understanding of the cutting-edge research and specific research-based challenges that are currently being faced in the MI field.

MI Course: Final Project (Week 16)

The final project is based on the results from homework 1, 2, and 3, where the assignments are tied together in implementation of MI on student's own data. Figure 4 illustrates the role of materials domain knowledge, data management skills, and machine learning skills for developing and implementing a MI project for a student's project on the assembly of nanoparticles [34]. Figure 4(a-c) illustrate the results from homework 1, 2, and 3 respectively. Figure 4(d) illustrates the student's final project workflow demonstrating the interplay between the assigned homework and how they form the fundamental steps to developing a performing a MI project.

The student's project required the implementation of data management skills, coding, and machine learning algorithms. Figure 4(b) illustrates the construction of a database schema which the student developed in homework 2. This was built off homework 1 which lists the valuable information of the nanoparticle system from standard characterization techniques. Through this process, the students should gain familiarity with reading, writing and parsing relevant common data file formats such as unstructured texts, CSV, and Microsoft Excel using one of the preferred programming languages or libraries. A knowledge of SQL query is also helpful to access large industry standard databases. SqLite3 databases are file based, portable, supported by libraries such as RSQlite in R, sqlite3 in python and built into MATLAB. Tools such as SQLiteBrowser can be used to easily import or export data from the database to CSV format and visualize and maintain the database. While handling the raw data for database development, students will note that a collected raw dataset often contains missing values, noise, errors and outliers. The next logical step is to understand the raw data and try to eliminate these limitations as much as possible before an ML algorithm can be applied. A collected dataset should go through the following refinements:

Downloaded from https://www.cambridge.org/core. NC State University Libraries, on 24 Jul 2020 at 17:42:12, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1557/adv.2020.171

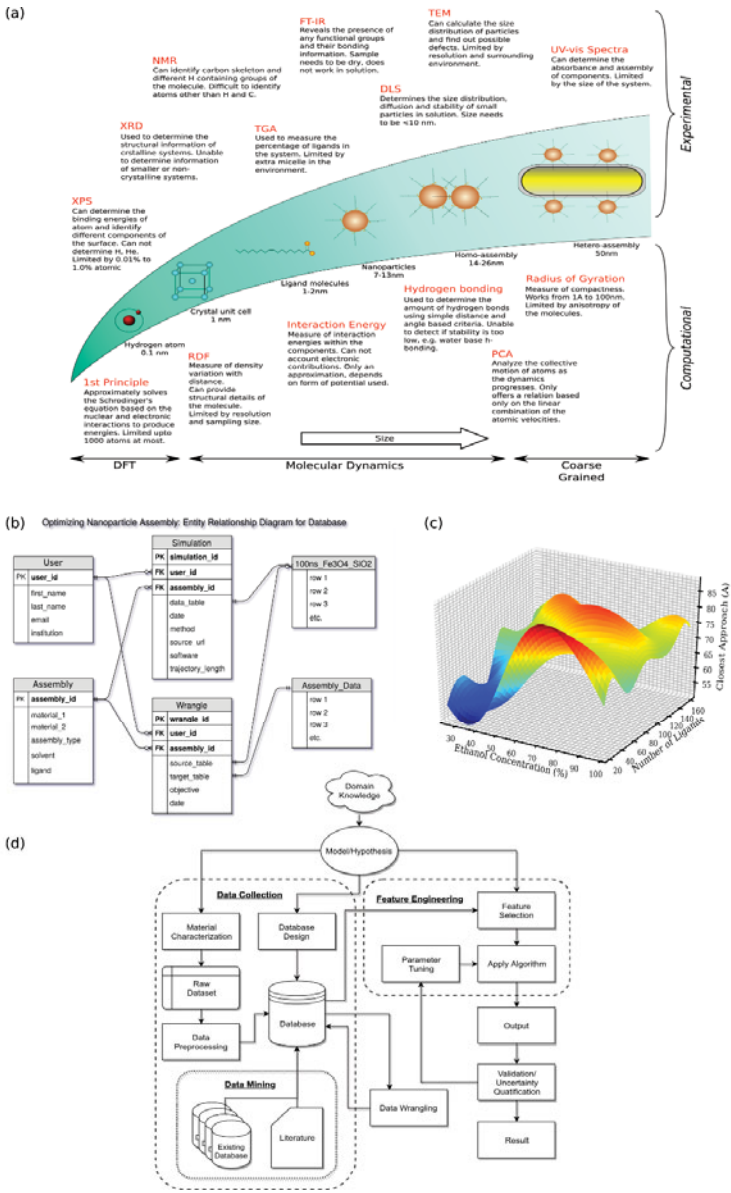

Figure 4. (a) the final project workflow based off (a) homework 1 (b) homework 2 and (c) homework 3. (d) highlights the overall workflow that is developed when combining homework 1, 2, and 3. This figure ultimately demonstrates the homework teach students how to develop and perform a MI project in manageable increments. <<color online>>

Downloaded from https://www.cambridge.org/core. NC State University Libraries, on 24 Jul 2020 at 17:42:12, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1557/adv.2020.171

1. Data Profiling: the student should try to understand the relation between the descriptors, their original input conditions, and consistency of the datasets if multiple sources were used. This step also involves connecting different raw dataset with the right descriptors and the hypothesis.

2. Data Structuring: using the available data analytics tools or programming languages, the raw data should be formatted into an accessible format that can be easily read by the ML model. While multiple recommendations have been proposed, such as XML, JSON, SQL, there is no consensus among the material science community about a standard format for data storage and management. We recommend storing the data in the tables of a relational database as a good practice. Relational databases are industry standard, easy to query, filter, and can be easily exported and imported to other common formats such as CSV. However, depending on the type of data, other options may include NoSQL databases, graph databases, time series databases etc. For smaller datasets, less technical alternatives such as MS Excel, CSV format, panda's dataframe etc. may be more suitable for the purpose of the project.

3. Data Cleaning: outliers and errors should be eliminated, noisy values should be smoothed out, inconsistent data should be corrected. One may need to come up with alternatives to the missing values by ad hoc means, for example, running experiment or simulation, regression, or averaging the available values. Common practice also involves using "Unknown" or "N/A" to denote the missing values.

Once the raw data and database development are understood, students must focus on manipulating the data for use. Data wrangling depends on the question under study. While there are no fixed rulesets for wrangling, it should have a clear set of objectives that will closely follow the hypothesis. Wrangling may involve dropping the null or unknown data or filtering and grouping specific features and descriptors. This process will often require the normalization or scaling of features to remove biases. Data wrangling can also include discretization, where a feature can be divided into multiple features, and aggregation, where multiple features are merged into a single feature. While a student will not receive hands on exposure to all wrangling processes, this course should expose the student to relevant data processing that can be successfully used in their respective project. Depending on the student's objective and hypothesis, multiple data wrangling can be done on the cleaned raw dataset. So, a good practice is to leave the raw data set alone

Downloaded from https://www.cambridge.org/core. NC State University Libraries, on 24 Jul 2020 at 17:42:12, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1557/adv.2020.171

for future use. As you can see in Figure 4(b), the student's database structure creates a new table that was used as a relational database by filtering and transforming the raw data table to generate a new wrangled data table.

To understand and verify some basic relationships within the data, visual analysis, graphing, plotting, correlation and covariance analysis should be performed. This will help the student get a general idea of the working dataset and find any patterns that may become useful later. The next step is to select the correct features that can be input to an ML algorithm. While choosing the correct descriptors depends on the hypothesis, feature importance and feature selection can often be automated by algorithms that are included in most data analytics tools, such as scikit-learn. These algorithms can evaluate the importance of each feature by statistical testing and the ability of each feature to make accurate predictions after being trained. These feature selection algorithms are broadly divided into 3 types, namely, filtering method, wrapper method, and embedded method. Validation techniques are the final step in feature selection to make sure the training data and descriptive features do not lead to overfitting or selection bias.

After cross validation, when the algorithm shows good accuracy with the testing data, the model is then ready for prediction. In the example shown in Figure 4, a Gaussian Process was trained to create the predictive model. Material science domain knowledge is necessary to evaluate the prediction made by the model, but as this student learned, the MI approach has other built in benefits. The Gaussian Process creates its predictions based on non-parametric fitting to the data it is trained on. This trained model is uniquely suited for extrapolation and interpolation with the power to quantify the uncertainty in the predicted values. A high uncertainty prompted the student to conduct more experiments in the uncertain area which was fed back into the model to increase prediction accuracy. Thus, outside of predictive analysis, the student saw the value in using MI for experimental design.

MI Course: Summary

Overall, the goal of the Materials Informatics course was to introduce the emerging field of materials informatics along with current approaches that employ machine learning to accelerate the process of materials optimization, discovery, and development compared to traditional experiments or computations. This goal was accomplished by hitting a series

Downloaded from https://www.cambridge.org/core. NC State University Libraries, on 24 Jul 2020 at 17:42:12, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1557/adv.2020.171

of student learning outcomes. At the end of the course, the students should be able to:

1. Describe the types of machine learning and understand how materials database function
2. Demonstrate an understanding of key materials informatics concepts and components
3. Demonstrate an understanding of supervised learning algorithms and identify materials problems that can be addressed using these techniques
4. Demonstrate an understanding of unsupervised learning algorithms and identify materials problems that can be addressed using these techniques
5. Identify algorithms that can be used for optimization problems in materials research
6. Evaluate existing and emerging machine learning technologies and analyze trends in data-driven techniques to anticipate how materials informatics evolve to meet changing need

MI COURSE STUDENT ASSESSMENT

Student assessment before and after the course indicated that upon completion of the course in this format, it is evident that students improved their understanding and confidence for implementing materials informatics concepts in their own research, as seen in Figure 5. As discussed earlier,

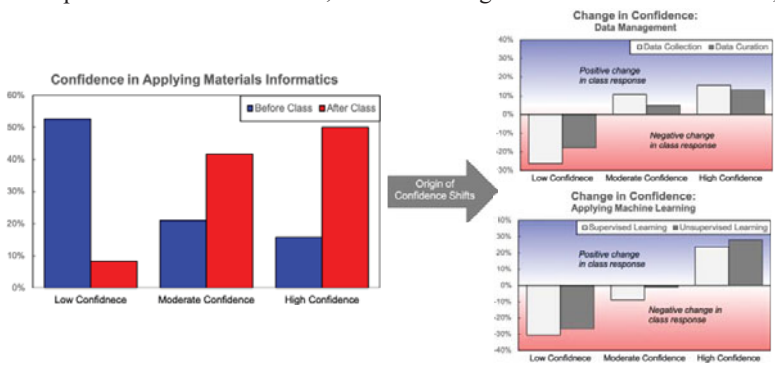

Figure 5. (a) Illustrates an increase in student confidence for applying materials informatics and (b) breaks down the increase of overall confidence into data management and machine learning categories. <<color online>>

Downloaded from https://www.cambridge.org/core. NC State University Libraries, on 24 Jul 2020 at 17:42:12, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1557/adv.2020.171

students must possess a specific skillset to be confident in using materials informatics, which included data management for collecting and pre-processing information as well as machine learning algorithms for understanding the informatics results and processes. Figure 5(a) highlights the overall rise in student confidence for applying materials informatics in their own research while Figure 5(b) looks at the potential skillsets that served as the driving force for this overall change. Specifically, it appears that the largest progress revolved around an increase in confidence for utilizing the machine learning algorithms. This large increase likely arises from the nature of the students taking the course. The class was entirely composed of PhD students in STEM fields; thus, data collection and curation are not as foreign to most students as compared to the machine learning algorithms. The results from the assessment indicates that the course structure can be a valuable template for other universities to implement in their own materials science curriculum.

CLOSING STATEMENT

As MI grows in academia and industry there will be a significant need for qualified students to fill labor demands at research institutions and companies. The course layout described above provides the blueprints for universities to include MI in their curriculum, ultimately preparing the next generation of students to enter the workforce with the necessary skillset for MI. Outside of the classroom, there are 'low hanging fruit' research opportunities to jump-start students interested in MI. This includes opportunities for students to repurpose the established methodology for different materials and models. Repurposing methodology may be considered incremental research, but it provides meaningful results and can be a perfect entry point for students and faculty interested in contributing to the flourishing MI field.

ACKNOWLEDGMENTS

The authors acknowledge the funding provided by the National Science Foundation(CMMI-1727603 and CMMI-1763025) and the NSF Research Traineeship on Data-Enabled Science and Engineering of Atomic Structures (DGE-1633587).

Downloaded from https://www.cambridge.org/core. NC State University Libraries, on 24 Jul 2020 at 17:42:12, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1557/adv.2020.171

References:

[1] H. Chan *et al.*, "Machine learning coarse grained models for water," *Nat. Commun.*, 2019.

[2] C.-T. Chen and G. X. Gu, "Composite Materials: Effect of Constituent Materials on Composite Performance: Exploring Design Strategies via Machine Learning (Adv. Theory Simul. 6/2019)," *Adv. Theory Simulations*, vol. 2, no. 6, 2019.

[3] J. Behler, "Perspective: Machine learning potentials for atomistic simulations," *J. Chem. Phys.*, vol. 145, no. 17, 2016.

[4] J. Hill, G. Mulholland, K. Persson, R. Seshadri, C. Wolverton, and B. Meredig, "Materials science with large-scale data and informatics: Unlocking new opportunities," *MRS Bull.*, vol. 41, no. 5, pp. 399–409, 2016.

[5] S. Curtarolo, G. L. W. Hart, M. B. Nardelli, N. Mingo, S. Sanvito, and O. Levy, "The high-throughput highway to computational materials design," *Nat. Mater.*, pp. 191–201, 2013.

[6] Y. Liu, T. Zhao, W. Ju, S. Shi, S. Shi, and S. Shi, "Materials discovery and design using machine learning," *J. Mater.*, vol. 3, no. 3, pp. 159–177, 2017.

[7] K. Takahashi and Y. Tanaka, "Material synthesis and design from first principle calculations and machine learning," *Comput. Mater. Sci.*, vol. 112, pp. 364–367, 2016.

[8] L. R. Zhao, K. Chen, Q. Yang, J. R. Rodgers, and S. H. Chiou, "Materials informatics for the design of novel coatings," *Surf. Coatings Technol.*, vol. 200, no. 5–6, pp. 1595–1599, 2005.

[9] S. Zeng, G. Li, Y. Zhao, R. Wang, and J. Ni, "Machine Learning-Aided Design of Materials with Target Elastic Properties," *J. Phys. Chem. C*, vol. 123, no. 8, pp. 5042–5047, 2019.

[10] R. Liu, A. Kumar, Z. Chen, A. Agrawal, V. Sundararaghavan, and A. Choudhary, "A predictive machine learning approach for microstructure optimization and materials design," *Sci. Rep.*, vol. 10, no. 1, 2015.

[11] S. Srinivasan *et al.*, "Mapping Chemical Selection Pathways for Designing Multicomponent Alloys: An informatics framework for materials design," *Sci. Rep.*, 2015.

[12] H. J. Kulik, "Making machine learning a useful tool in the accelerated discovery of transition metal complexes," *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 2019.

[13] C. Kim, G. Pilania, and R. Ramprasad, "Machine Learning Assisted Predictions of Intrinsic Dielectric Breakdown Strength of ABX3 Perovskites," *J. Phys. Chem. C*, vol. 120, no. 27, pp. 14575–14580, 2016.

[14] H. Nakata and S. Bai, "Development of a new parameter optimization scheme for a reactive force field based on a machine learning approach," *J. Comput. Chem.*, vol. 40, no. 23, pp. 2000–2012, 2019.

[15] P. Wang, Y. Shao, H. Wang, and W. Yang, "Accurate interatomic force field for molecular dynamics simulation by hybridizing classical and machine learning potentials," *Extrem. Mech. Lett.*, vol. 24, pp. 1–5, 2018.

[16] C. Chen, Z. Deng, R. Tran, H. Tang, I. H. Chu, and S. P. Ong, "Accurate force field for molybdenum by machine learning large materials data," *Phys. Rev. Mater.*, vol. 1, no. 4, 2017.

[17] V. Botu and R. Ramprasad, "Learning scheme to predict atomic forces and accelerate materials simulations," *Phys. Rev. B - Condens. Matter Mater. Phys.*, vol. 92, no. 9, 2015.

[18] M. A. Wood, M. A. Cusentino, B. D. Wirth, and A. P. Thompson, "Data-driven material models for atomistic simulation," *Phys. Rev. B*, vol. 99, no. 18, 2019.

[19] P. Bleiziffer, K. Schaller, and S. Riniker, "Machine Learning of Partial Charges Derived from High-Quality Quantum-Mechanical Calculations," *J. Chem. Inf. Model.*, vol. 58, no. 3, pp. 579–590, 2018.

[20] S. Chmiela, H. E. Sauceda, K. R. Müller, and A. Tkatchenko, "Towards exact molecular dynamics simulations with machine-learned force fields," *Nat. Commun.*, 2018.

[21] Y. Li *et al.*, "Machine Learning Force Field Parameters from Ab Initio Data," *J. Chem. Theory Comput.*, vol. 13, no. 9, pp. 4492–4503, 2017.

[22] T. D. Huan, R. Batra, J. Chapman, S. Krishnan, L. Chen, and R. Ramprasad, "A universal strategy for the creation of machine learning-based atomistic force fields," *npj Comput. Mater.*, 2017.

[23] P. Miles, L. Leon, R. C. Smith, and W. S. Oates, "Analysis of a multi-axial quantum informed ferroelectric continuum model: Part 1—uncertainty quantification," *J. Intell. Mater. Syst. Struct.*, vol. 29, no. 13, pp. 2823–2839, 2018.

[24] L. Leon, R. C. Smith, W. S. Oates, and P. Miles, "Analysis of a multi-axial quantum-informed ferroelectric continuum model: Part 2—sensitivity analysis," *J. Intell. Mater. Syst. Struct.*, vol. 29, no. 13, pp. 2840–2860, 2018.

[25] A. R. Paterson, B. J. Reich, R. C. Smith, A. G. Wilson, and J. L. Jones, "Bayesian approaches to uncertainty quantification and structure refinement from X-ray diffraction," in *Springer Series in Materials Science*, 2018, pp. 81–102.

Downloaded from https://www.cambridge.org/core. NC State University Libraries, on 24 Jul 2020 at 17:42:12, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1557/adv.2020.171

[26] W. Xu and J. M. LeBeau, "A Convolutional Neural Network Approach to Thickness Determination using Position Averaged Convergent Beam Electron Diffraction," *Microsc. Microanal.*, vol. 23, 2017.
[27] Louis Columbus, "Roundup Of Machine Learning Forecasts And Market Estimates, 2018," *Forbes*, 2018. [Online]. Available: https://www.forbes.com/sites/louiscolumbus/2018/02/18/roundup-of-machine-learning-forecasts-and-market-estimates-2018/#2c05d4602225. [Accessed: 10-Dec-2019].
[28] "Citrine Informatics," 2019. [Online]. Available: https://www.linkedin.com/company/citrine-informatics/insights/. [Accessed: 12-Dec-2019].
[29] Kumaresh Pattabiraman, "LinkedIn's Most Promising Jobs of 2019," 2019. [Online]. Available: https://blog.linkedin.com/2019/january/10/linkedins-most-promising-jobs-of-2019. [Accessed: 12-Dec-2019].
[30] "Mathematicians and Statisticians," *Occupational Outlook Handbook*, 2019. [Online]. Available: https://www.bls.gov/ooh/math/mathematicians-and-statisticians.htm. [Accessed: 12-Dec-2019].
[31] Linda Burtch, "The Burtch Works Study Salaries of Data Scientists & Predictive Analytics Professionals," 2019.
[32] V. Venkatraman and B. Alsberg, "Designing High-Refractive Index Polymers Using Materials Informatics," *Polymers (Basel).*, 2018.
[33] J. S. Peerless, N. J. B. Milliken, T. J. Oweida, M. D. Manning, and Y. G. Yingling, "Soft Matter Informatics: Current Progress and Challenges," *Adv. Theory Simulations*, vol. 2, no. 1, 2019.
[34] M. D. Manning, A. L. Kwansa, T. Oweida, J. S. Peerless, A. Singh, and Y. G. Yingling, "Progress in ligand design for monolayer-protected nanoparticles for nanobio interfaces," *Biointerphases*, vol. 13, no. 6, 2018.
[35] J. A. Nash, A. L. Kwansa, J. S. Peerless, H. S. Kim, and Y. G. Yingling, "Advances in molecular modeling of nanoparticle-nucleic acid interfaces," *Bioconjug. Chem.*, vol. 28, no. 1, pp. 3–10, 2017.
[36] N. K. Li *et al.*, "Prediction of solvent-induced morphological changes of polyelectrolyte diblock copolymer micelles," *Soft Matter*, vol. 11, no. 42, pp. 8236–8245, 2015.
[37] D. Weininger, "SMILES, a Chemical Language and Information System: 1: Introduction to Methodology and Encoding Rules," *J. Chem. Inf. Comput. Sci.*, vol. 28, no. 1, pp. 31–36, 1988.
[38] D. Weininger, A. Weininger, and J. L. Weininger, "SMILES. 2. Algorithm for Generation of Unique SMILES Notation," *J. Chem. Inf. Comput. Sci.*, vol. 29, no. 2, pp. 97–101, 1989.
[39] T.-S. Lin *et al.*, "BigSMILES: A Structurally-Based Line Notation for Describing Macromolecules," *ACS Cent. Sci.*, vol. 5, no. 9, pp. 1523–1531, 2019.
[40] E. De Guire *et al.*, "Data-driven glass/ceramic science research: Insights from the glass and ceramic and data science/informatics communities," *J. Am. Ceram. Soc.*, vol. 102, no. 11, pp. 6385–6406, 2019.
[41] O. Kononova *et al.*, "Text-mined dataset of inorganic materials synthesis recipes," *Sci. data*, 2019.
[42] H. M. Berman *et al.*, "The Protein Data Bank (www.rcsb.org)," *Nucleic Acids Res.*, 2000.
[43] F. C. Bernstein *et al.*, "The Protein Data Bank," *Eur. J. Biochem.*, vol. 80, no. 2, pp. 319–324, Nov. 1977.
[44] S. K. Burley *et al.*, "RCSB Protein Data Bank: Biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy," *Nucleic Acids Res.*, vol. 47, pp. D464–D474, 2019.
[45] "Source: National Institute for Materials Science." [Online]. Available: https://www.nims.go.jp/eng/. [Accessed: 09-Dec-2019].
[46] P. Villars *et al.*, "The Pauling File, Binaries Edition," in *Journal of Alloys and Compounds*, 2004.
[47] S. Otsuka, I. Kuwajima, J. Hosoya, Y. Xu, and M. Yamazaki, "PoLyInfo: Polymer database for polymeric materials design," in *Proceedings - 2011 International Conference on Emerging Intelligent Data and Web Technologies, EIDWT 2011*, 2011.
[48] K. Anderson *et al.*, "Creating the Next Generation Materials Genome Initiative Workforce," 2019.
[49] R. Mansbach *et al.*, "Reforming an undergraduate materials science curriculum with computational modules," *J Mater Educ*, vol. 38, pp. 161–174, 2016.
[50] "Data-Enabled Science and Engineering of Atomic Structure (SEAS)." [Online]. Available: https://www.mse.ncsu.edu/seas/traineeship/. [Accessed: 16-Dec-2019].
[51] W. Li, R. Jacobs, and D. Morgan, "Predicting the thermodynamic stability of perovskite oxides using machine learning models," *Comput. Mater. Sci.*, vol. 150, pp. 454–463, 2018.

Downloaded from https://www.cambridge.org/core. NC State University Libraries, on 24 Jul 2020 at 17:42:12, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1557/adv.2020.171