PRINCETON UNIVERSITY

DEPARTMENT OF CHEMICAL AND BIOLOGICAL ENGINEERING

CBE 512: *Machine Learning in Chemical Science and Engineering* Fall 2024

Last revised: *August 30, 2024*

==========

Lecture: Sherrerd Hall 101

T: 1:30-2:50pm (Eastern)

Th: 1:30-2:50pm (Eastern)

Instructor

Prof. Michael A. Webb

mawebb@princeton.edu

Office Hours: Th, 3:00-5:00pm, A325

Co-Instructor

Dr. Shengli (Bruce) Jiang

sj0161@princeton.edu

Office Hours: TBA

==========

Scope of the Course: This course provides a combined theoretical and practical introduction to machine learning methods and their application in chemistry, chemical engineering, and materials science. Following a broad survey of current machine learning algorithms, including supervised and unsupervised learning methods, as well as best practices, we delve into specific application areas (e.g., *quantitative structure-property relationship modeling, materials design, molecular simulation, process control, compound synthesis*) to ascertain how machine learning methods, which are commonplace in Big Tech, are applied in chemical/biological/materials research and engineering. Students will also gain exposure to state-of-the-art applications of machine learning in science and engineering through topical literature reviews and case studies. Course assignments are constructed to develop both understanding and proficiency with commonly used algorithms while deployed in a science/engineering context. The course also favors the deployment of existing algorithms using professional machine learning libraries with guided instruction with some limited *ab initio* algorithm implementation.

Rationale for the course: Data science and machine learning are increasingly important in the domain of "physical" scientists and engineers whose education and training is traditionally rooted in the physics and chemistry of molecules and/or materials. There are now needs in both academic and industrial settings for individuals to have core competencies in machine learning, while being a domain expert. Nevertheless, significant course instruction in ML may be theory-driven or abstract in that it is far from expected domains of application for a physical engineer. This course addresses a content gap between the traditional presentation of machine learning by premising content/instruction/examples in the domain of chemistry and materials. This is particularly manifest in datasets based on materials property prediction, design, or characterization that position algorithms/application closer to examples that are likely to be encountered by, e.g., a materials scientist, chemist, or biological engineer. This approach also allows us to discuss and present strategies related to data scarcity, data imperfection/noise, and model assessment in the context of relevant data. Relative to other courses, the course provides broad topical exposure across many disciplines of machine learning, equipping one with the tools to engage in more detailed study in many areas, sometimes at the expense of more robust theoretical development.

Teaching Objectives and Outcomes for this Course:

1. This class will teach you fundamental guiding principles underlying common machine learning algorithms and strategies. This class will expose students to language and terms used in the context of machine learning. Students will be able to capably describe different branches of machine learning and provide distinguishing examples for their application and goals for their implementation.

2. This class will showcase applications in contexts that are broadly more relevant to "physical" engineers and scientists. Students will be able to more confidently approach and analyze topical literature.

3. This class will expose you to particular approaches and strategies for representing molecules and materials in the context of machine learning. Students will be able to formulate machine learning problems in the form of structure-property relationships of interest to physical scientists and engineers.

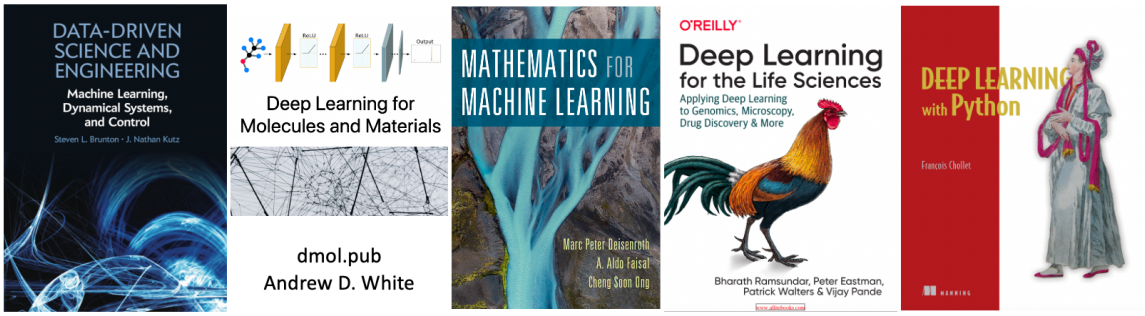
4. The class will give you practical experience and knowledge with how to interface with software common for data science. Students will be able to prototype, debug, and test codes for data science applications. Students will have computational tools and practical knowledge needed to pursue more advanced topics in machine learning and understand their function.

Management: Canvas will be the official management tool for the class. On Canvas, you will find all course material including announcements, reading, assignments, etc.
https://princeton.instructure.com/courses/15679

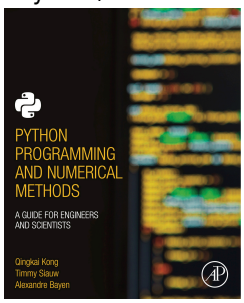
GitHub will also be used as a supplemental resource to host assignments/data/materials.
https://github.com/webbtheosim/CBE512-MLinChmSciEng

Chat/Discussion: In addition to typical course interactions (lecture, office hours), we will host a Slack channel for real-time interactions amongst students/instructors. Course announcements typically will be made on both Canvas and on Slack.
https://join.slack.com/t/cbe512/shared_invite/zt-2pudno94k-~EUoUK~DGmf2_mlnbSCvIQ

Course Materials: There is no required a textbook. The reality is that no single text book successfully synthesizes all the information, in a digestible format, covered in this course. You are expected to perform additional reading/research based on your own needs and interests. For machine learning, there is no shortage of content out of the web, and I have learned a great deal from kind internet souls. Selected readings will be identified and posted to the course website as guides. Additional helpful content (e.g., online tutorials, documentation pages, etc.) will also be referenced and linked. The combination of course notes, reading material posted to the course website, web searches, and documentation pages online should be sufficient for understanding course content at the intended level and completing coursework. Nonetheless, there is value in dedicated reading of a more formal resource. Some good references that we may use during the course are shown below.



Should you need some help/introduction with Python, this resource is hosted online for free:



Course Evaluation: Performance in class will be evaluated based on a combination of factors as shown below. Pending additional changes, the grading breakdown will be:

- **Problem Sets/Programming Assignments:** 40%
- **Midterm Paper & Presentation:** 15%
- **Class Participation/Presentation:** 15%
- **Final Project/Paper/Presentation:** 30%

There will be five problem sets/programming assignments distributed throughout the semester; these are distributed roughly once every 2-2.5 weeks. The purpose of these assignments is to solidify core principles covered in lecture, mostly through the implementation and extension of introduced algorithms/methods. You will be required to turn preferably turn in a python notebook or otherwise a sheet summarizing results with brief discussion as well as relevant scripts used to generate the results. In lieu of a midterm exam, there will be a midterm paper analyzing the algorithms and results presented in a research article as approved by the course

instructor. In lieu of a final exam, there will be a final project and report on a machine-learning problem chosen in conjunction with course instructors. Both the midterm and the final feature a presentation component. In addition, some lectures may have discussion surrounding assigned reading. The points allocated to participation/presentation refer to these latter elements.

Course Collaboration Policy: Discussion on course material and problem sets, with other classmates as well as the instructor and AIs, is not only permitted but highly encouraged using the tools available. The course is not a competition, and we want everyone to succeed! However, all submitted solutions *must* represent *your own* work. In addition, students are encouraged to attempt problems on their own and well ahead of the due date prior to any significant discussion. If you collaborate extensively, please list the names of the students with whom you collaborated on your submission of the assignment. If you significantly reference a resource outside of those provided in class, then you should indicate such to avoid the appearance of presenting someone else's work as your own. This also applies to the utilization of ChatGPT or related large-language models in facilitating any coursework, if significant.

Course Content: Lecture periods aim to provide a formal framework for exposure to algorithms, after which the instructor/class will usually engage in hands-on implementation/demonstration/evaluation of algorithms using domain-relevant examples. The expectation is that this structure, in combination with reading and perhaps a little help (from instructors, classmates or Google), will allow you to tackle more complex problems that appear on assignments, understand material in research papers, or even formulate and test problems in the setting of the student's research. Students should aim to preview any assigned reading material prior to attending lecture to get a flavor of the topics and note confusing elements. Reviewing the content after the lecture will help solidify the material. At times, lectures may cover material in more (or less) depth than that provided in the assigned reading. Questions on course content are welcome during lecture, office hours, and the course discussion venues. Lecture notes are usually posted slightly preceding lecture.

Tentative Course Schedule

*Target Due Date of Assignments

Date	*Topic(s)*
Week 01 *09/03/24* *09/05/24*	**Course Introduction:** data science in scientific discovery, materials synthesis & design; a brief history of ML and chemical science; course logistics; discussion on programming principles and software; environment setup **Math Review:** overview of mathematics and areas of application; key concepts from linear algebra; key concepts from analytic geometry; key concepts from vector calculus; key concepts in probability
Week 02 *09/10/24* *09/12/24**	**Overview of Machine Learning:** basic conceptualization and terminology in ML; supervised learning (regression and classification, examples on conductivity, melting point, band gaps, solubility, drug activity synthesis); unsupervised learning (clustering of chemical classes, signal processing and protein folding, identification of molecular states); reinforcement learning (process synthesis and control, nanofabrication); semi-supervised and self-supervised (microscopy segmentation and classification) **Intro to regression:** basic problem of curve-fitting and loss functions; linear regression and non-linear regression; gradient descent (vanilla, with momentum, stochastic); Activity: parameter optimization and regression with dataset on intrinsically disordered proteins and models informed by polymer physics. (numerical implementation, testing hypotheses, and comparing to linear algebraic solutions) **Breaking Down Engineering Problems for ML**: overview of common types of chemical data (spectra, time-series, images, molecular structures), common data structures, transformations, and modeling; deconstruction of case-studies in terms of ML vernacular on endotoxin detection, contaminant detection, electrochemistry, plastics recycling, molecular engineering, and medical imaging
Week 03 *09/17/24* *09/19/24*	**Polishing regression:** distinction of ML from linear and non-linear least-squares regression; multivariate regression and example (cost evaluations on unit operations and process equipment); generalization of loss **Feature scaling and data transformations:** motivation of feature scaling; explanation of common approaches (min-max, standard-scaling, max-abs, robust transformer, quantile transformer, power transforms, norm scaling); discussion of effects on data distribution; examples of scaled chemical data and considerations; demonstrations of transforms in scikit-learn **Model Selection I:** definitions and discussion of parameters vs. hyperparameters; over-determined vs. under-determined systems; underfitting vs. overfitting; bias-variance tradeoff; purpose, approach, and discussion on train-validation-test splits; cross-validation and common manifestations (k-fold, stratified, LOO, nested); deconstruction of nested k-fold from protein design paper; stratified selection according to polymer architectures; introduction to sklearn.model_selection functions; Activity: creating train/test splits and cross-validation folds from melting temperature dataset).
Week 04 *09/24/24* *09/26/24**	**Model Selection II:** Evaluating model complexity (Pareto optimality, Occam's razor); tenets of regularization and manifestation in molecular model force-field optimization; norm-based regularization; Activity: understanding L1 vs. L2 regularization and their effects with random linear-equations and visualization; discriminating models with information criteria (e.g., Bayesian information criterion, Akaike information criterion); connection between maximum likelihood estimation and least-squares minimization; feature selection strategies; overview of hyperparameter tuning and grid-search **Diving into classification I:** examples of classification tasks (phase behavior, toxicity); logistic regression; discussion of binary classification task; extension to multi-class problems (one vs. one and one vs. rest models) **Diving into classification II:** key concepts of support vector machine; different formulation of SVM and associated hyperparameters; visualization of decision boundaries for different hyperparameters; extension to multi-class; basics of decision trees and splitting strategies; impurity functions and entropy; CART algorithm and stopping criteria; hyperparameters of decision trees; visualization of decision trees and tree rules; explanation of random forest and ensembling; discussion of hyperparameters; Activity: comparison of random forest and SVM with different hyperparameters for molecule classification tasks

Week 05 *10/01/24* *10/03/24*	**Introduction to Neural Networks:** origin and basic structure of artificial neural networks; terminology; examination of single-layer perceptron vs. linear regression; examining non-linear activation; deconstruction of a neuron; Activity: generation of logic operations and resolution of XOR; universal approximation theorem; graphical intuition underlying universal approximation theorem; backpropagation and autodifferentiation **Introduction to Keras:** discussion of different API and utilization; overview of steps to creating neural networks; demonstration using sequential model; equivalent demonstration with functional model API; illustration of complex model with functional API and discussion of architecture; discussion of compile/fit method along with options/specifications; revisiting optimizers, explanation of NADAM and comparison of optimizer performance; discussion of model metrics within API and implementation of custom metrics/losses; Activity: examination of data and regression using neural networks for chemical solubility dataset
Week 06 *10/08/24* *10/10/24**	**Molecular Featurization I:** the necessity of featurization for QSPR and other modeling tasks in chemical science and engineering; discussion of what constitutes a molecule; principles of molecules in terms of graphs; motivation and discussion of text-based representations to convey molecular graphs; description of Simplified Molecular-Input Line Entry System and algorithm; description of SMILES Arbitrary Targe Specification for chemical pattern representation; Wiswesser line notation, SYBYL line notation, INChI codes; motivation and discussion of Self-referencing embedded strings (SELFIES) **Molecular Featurization II:** Ideas underlying tokenization in natural language processing; examination of words, n-grams, characters in the context of bioinformatics and biopolymer representations; discussion of one-hot-encoding; examples of one-hot-encoding and comparison to other strategies for polymer property prediction; Activity: developing one-hot representations of small-molecules from SMILES string characters
Week 07 *10/15/24* *10/17/24*	Fall Break Fall Break
Week 08 *10/22/24* *10/24/24*	**Molecular Featurization III:** presentation of extended connectivity fingerprints (ECFPs) as molecular representations; discussion of algorithm through self-consistent topological hashing; interpretation through one-hots; discussion of advantages/limitations of ECFPs; discussion of geometric representations; considerations of equivariance and invariance and physical properties of molecules; examination of various structural/geometric representations such as Coulomb Matrix, atom-centered symmetry functions; discussion of descriptor vectors and common physiochemical descriptors; case studies of molecular descriptors for solvent selection in catalysis, thermal transport in porous media, methane adsorption in nanoporous materials, and biological activities of nanoparticle composites **Workshop on descriptor libraries and software:** utilization and exploration of RDKit; utilization and exploration of Mordred; discussion of DScribe
Week 09 *10/29/24** *10/31/24*	Class Presentations on literature paper Class Presentations on literature paper

Week 10 *11/05/24* *11/07/24*	***Data clustering and unsupervised learning:*** *motivation for clustering; methods for defining similarity, including notions of molecular/chemical similarity via common metrics (e.g., Tanimoto, Earth Mover's distance); common issues with clustering and data conditioning/preprocessing; applications of clustering algorithms in charge transport and coarse-grained modeling; discussion of different categories of clustering approaches with representative methods (e.g., k-means vs. Gaussian mixtures vs. DIANA vs. DBSCAN); examination of k-means; implementation in scikit-learn; Activity: comparison of clustering over copolymer formulation space* ***Dimensionality reduction and unsupervised learning:*** *motivation for dimensionality reduction; different categories/approaches; discussion of principal component analysis with example over chemical dataset; illustration of implementation; discussion of manifold-learning and isomap algorithm; case-studies of dimensionality reduction in the materials literature such as use of variational autoencoders in conjugated peptides, analysis of 19F MRI agents from copolymer synthesis, discovery of high-temperature polymers; characterization of protein-folding; discussion of implementations and restrictions in scikit-learn; Activity: comparison of dimensionality reduction techniques (PCA, isomap, t-SNE) for processing a molecular dynamics trajectory of alanine dipeptide*
Week 11 *11/12/24** *11/14/24*	**Interrogating "black-box" models:** self-explaining models; attribution methods; counterfactual explanations; discussion of applications to chemical spaces/explanations (enzyme activity, biocondensate formation, blood-brain barrier permeation, solubility prediction, scent-structure relationships) **Examination of "Best Practices":** discussion of domain resources; FAIR data principles; discussion of common materials data repositories; important considerations regarding sourcing and reporting data; considerations and strategies regarding data sanitization; special considerations regarding data transformation and scaling; tools for hyperparameter optimization; utilization of baseline simple models; revisitation of overfitting with assessment and mitigation strategies
Week 12 *11/19/24* *11/21/24*	**Gaussian process regression:** overview and key points; intuition for Gaussian processes through construction of Gaussian vectors with added covariance; theoretical underpinnings of Gaussian processes and basic algorithm; Activity: notebook demonstration of implementation **Data seeding and design of experiments:** ex nihilo data generation and materials design; the curse of dimensionality and motivation for effective seeding; common strategies and essentials of algorithmic implementation; implications and comparison against Bayesian optimization; Activity: charting out phase behavior of an active matter system **Active learning and Bayesian optimization:** motivation and areas of application in materials and formulation design; terminology and literature distinctions; discussion of active learning frameworks and objectives from the literature such as in peptide design, compound screening, therapeutic development, energy-storage materials, and biomaterials optimization; practical considerations/important elements to active learning structure; discussion and comparison of acquisition functions; discussion of robotic platforms and autonomous experimentation
Week 13 *11/26/24* *11/28/24*	Philosophical Debate or Special Topic Focus I Thanksgiving Recess
Week 14 *12/03/24** *12/05/24*	Special Topic Focus II/Overflow Special Topic Focus III/Overflow
Week 15 *12/12*	End-of-term Presentations/Reports

Special Topics may include: *advanced neural network architectures, transformers, diffusion models, large-language models in chemistry, grey-box models and physics-informed machine learning, etc.*
These will be selected by discussion/class vote