Developing a Machine Learning Course for Chemical Engineering

Alex Albaugh

AIChE National Meeting
October 4, 2025
Boston, MA

A Course For All ChE/MSE Students

- Wayne State Chemical Engineering & Materials Science
  - 26,000 student public R1 university in Detroit, MI
  - 91 B.S. students
  - 19 M.S. students
  - 18 PhD. students
  - 13 faculty

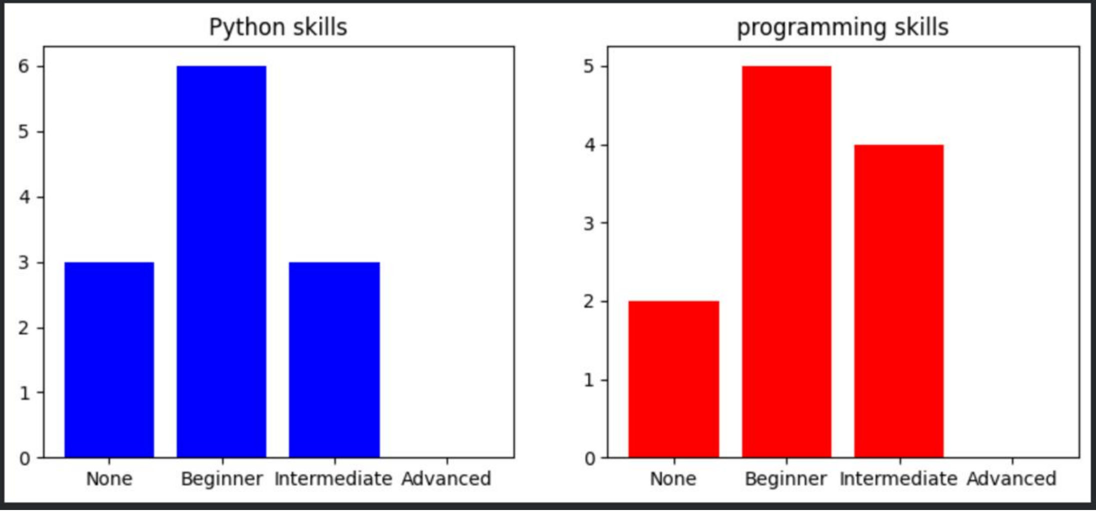
- Motivation
  - Modernize curriculum
  - Expand elective offerings
  - Integrate computational methods
- Designed for B.S., M.S., and Ph.D. students
  - Cross-listed at upper division undergrad/PhD level
  - In-person, shared lectures
  - Students from biology, chemistry, physics
  - Not a programming course!

Python, Our Language of Choice

WAYNE STATE
UNIVERSITY

- Used in intro. programming course
- Widely adopted
- Portable
- User friendly
- Libraries for everything
  - `numpy`: computation
  - `pandas`: data I/O
  - `matplotlib`: plotting
  - `sklearn`: general ML
  - `keras`: neural networks
- Free

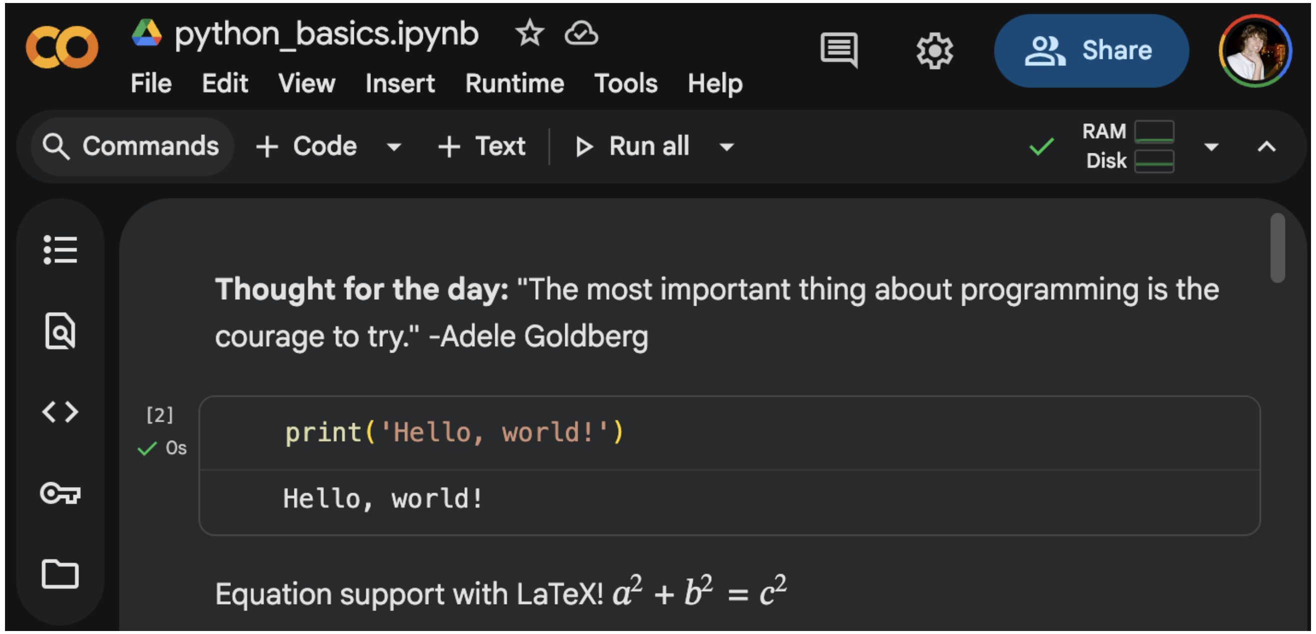
```python
import numpy as np
import matplotlib.pyplot as plt

categories = ['None','Beginner','Intermediate','Advanced']
python = [3,6,3,0]
programming = [2,5,4,0]

fig,axes = plt.subplots(1,2,figsize=(10,4))
axes[0].bar(categories,python,color='b')
axes[1].bar(categories,programming,color='r')
axes[0].set_title('Python skills')
axes[1].set_title('programming skills')

plt.show()
```

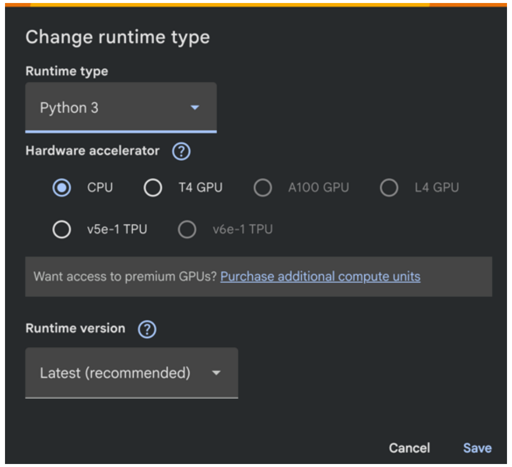


Lesson 2: Python Basics

Google Colab, Our Environment of Choice

- Supports notebooks (.ipynb)
- No installation
- No version management
- No environment management
- Compatible with any machine
- “Explain error” feature
- Requirements:
  - Browser + internet
  - Google account
- Server-side compute
- Access to advanced hardware
- Free





Lesson 2:
Python Basics

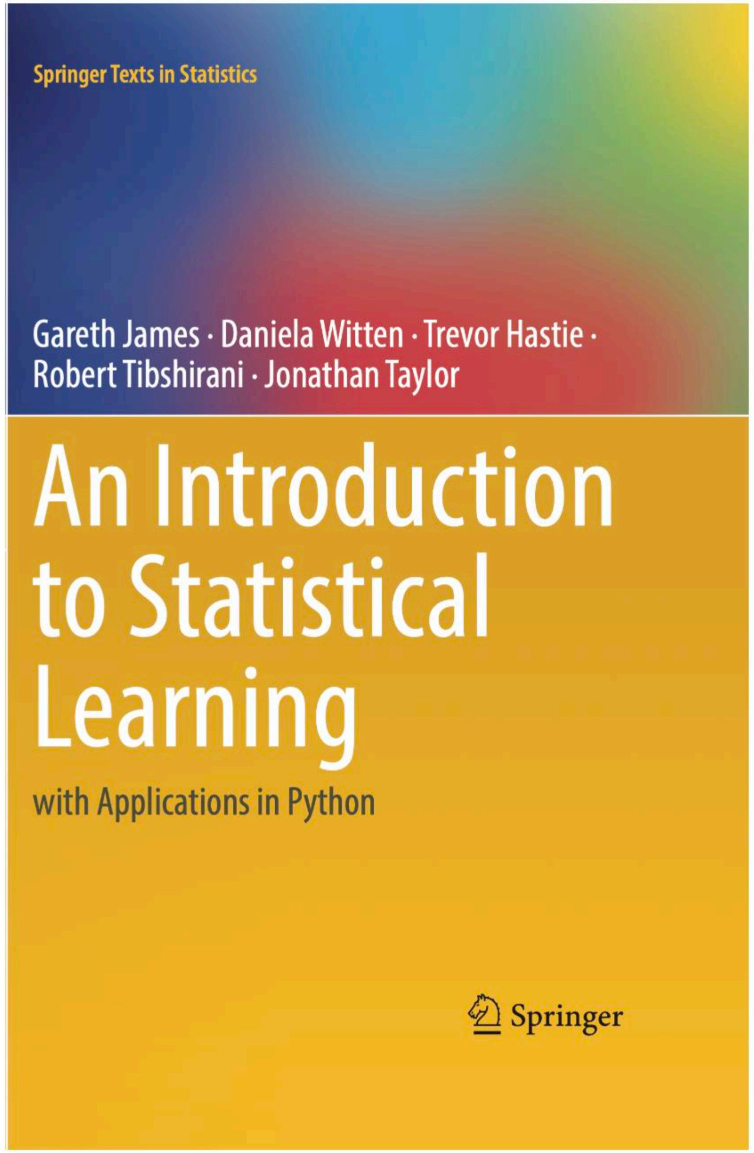
Tip: Share Data Sets via Github URL

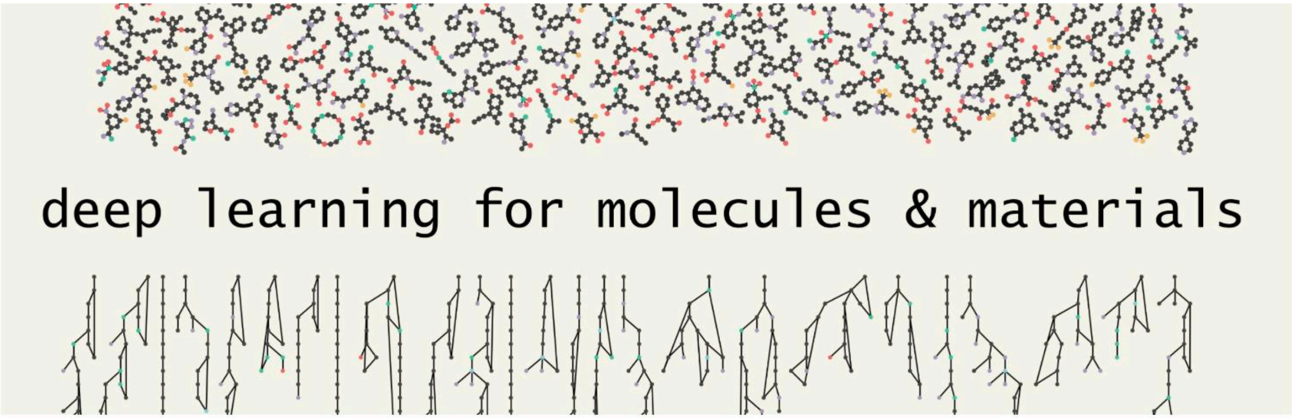
WAYNE STATE UNIVERSITY

- Upload data sets to public Github
- Share URL
- Load directly from URL in Colab
- No download/upload issues
- Free

```
[2]
✓ 0s
#read data
df = pd.read_csv('https://raw.githubusercontent.com/albaugh/CHE7507/refs/heads/main/Lecture8/combined_cycle_power_plant.csv')

#set features
X = df[['AT', 'AP', 'RH', 'V']].values

#set target
y = df['PE'].values
```

Bridging Intro. ML and ChE Applications





Andrew White

Great chemical engineering and chemistry examples

Jumps in with advanced methods

Free

Excellent introduction to concepts with broad coverage

No chemical applications

Free

statlearning.com, dmol.pub

Lesson Structure: Theory + Practice

Convolutions

A convolution applies a filter to an input to find patterns that match the filter. For a 1-D input f with a 1-D filter g of length K, the convolution is:

$$(f * g)(m) = \sum_{i=0}^{K-1} f(m-i)\,g(i)$$

Let's look at example. Here our input sequence is $f = [0, 0, 0, 10, 10, 10, 10]$ and our filter is $g = [-1, 1]$.

1st element

$$f = [0, 0, 0, 10, 10, 10, 10]$$
$$g = [-1, 1]$$
$$(f * g)(0) = 0(-1) + 0(1) = 0$$

2nd element

$$f = [0, 0, 0, 10, 10, 10, 10]$$
$$g = [-1, 1]$$
$$(f * g)(1) = 0(-1) + 0(1) = 0$$

2-D Convolutional Neural Network

2-D CNNs are useful for image classification. In an engineering setting they can be useful for real-time monitoring of prodcuts on an assembly line. Since they are typically used for image classification, I don't have a great example to show you. Training an image classifier in lecture is going to be prohibitively expensive. I can, however, show you the synthax for building a 2-D CNN. Below is the `Keras` synthax that would build the CNN example at the end of lecture.

[ ]

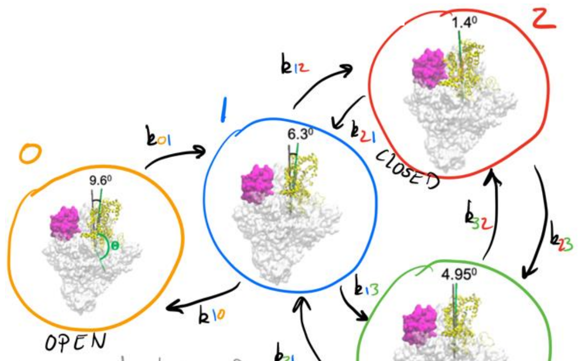
```
model = keras.Sequential([
  keras.layers.Input((32, 32, 3)),
  keras.layers.Conv2D(6, (3, 3), activation='relu', padding='same'), #convolution layer with 6 (3 x 3) filters and
  keras.layers.MaxPooling2D((2, 2)), #max pool layer with a 2 x 2 filter

  keras.layers.Conv2D(12, (3, 3), activation='relu', padding='same'), #convolution layer with 12 (3 x 3) filters an
  keras.layers.MaxPooling2D((2, 2)), #max pool layer with a 2 x 2 filter

  keras.layers.Conv2D(24, (3, 3), activation='relu', padding='same'), #convolution layer with 24 (3 x 3) filters an
  keras.layers.MaxPooling2D((2, 2)), #max pool layer with a 2 x 2 filter

  keras.layers.Flatten(), #flatten the convolution output from a series of matrices into a single 1-D vector
  keras.layers.Dense(500, activation='relu'), #feed the flattened convolutions through a dense layer with 500 neuro
  keras.layers.Dense(100, activation='softmax')  #for classifications, the final layer works best with a 'softmax'
])

model.summary()
```

Model: "sequential"

Layer (type)	Output Shape	Param #
conv2d (Conv2D)	(None, 32, 32, 6)	168
max_pooling2d (MaxPooling2D)	(None, 16, 16, 6)	0

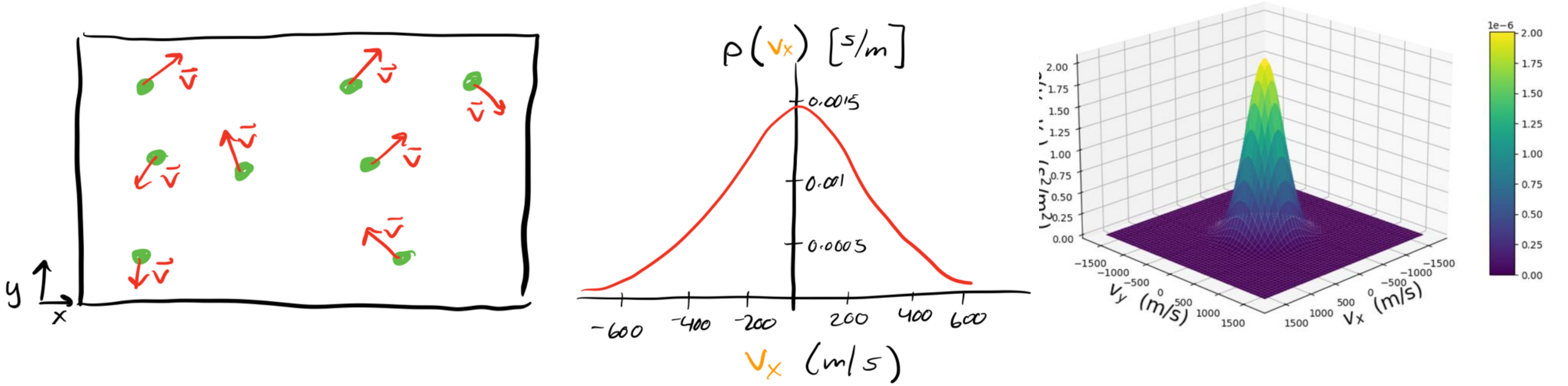
Content: Review of Building Blocks

Concept	Examples
Python programming, packages	
Linear algebra	Process modeling, Markov models, chemical reaction networks
Statistics	Kinetic theory of gases, Maxwell-Boltzmann statistics



Opening and closing of a bacterial RNA polymerase clamp, from "Markov State Models to Study the Functional Dynamics of Proteins in the Wake of Machine Learning," JACS Au, 2021, Konovalov, et al.

Lesson 4: Linear Algebra

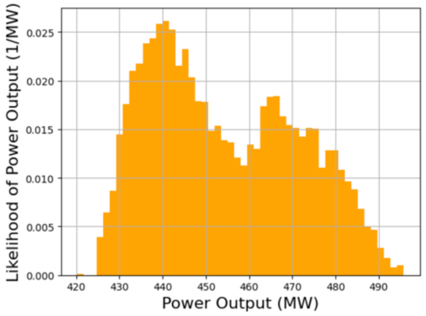


Lesson 21: Statistics

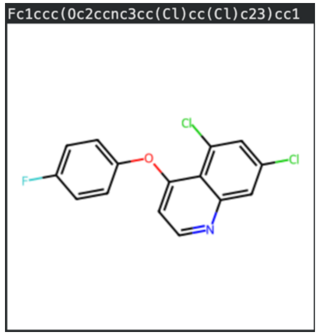
Konovalov, *et al.*, *JACS Au*, 2021

Content: Core Machine Learning Concepts

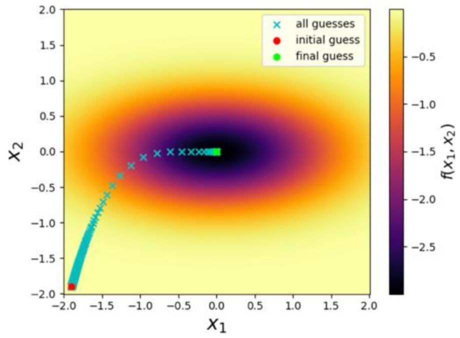
Concept	Examples
Regression, classification	Handwriting classification, FDA approval prediction
Bias-variance, training/testing/validating, regularization	Sabatier principle, power plant output prediction
Featurization	SMILEs strings, molecular descriptors, molecular features
Optimization	Optimizing reactor conditions
Ethics	LLMs & copyright law, ML in medical decision making, environmental impact



Lesson 8: Validation



Lesson 15: Featurization

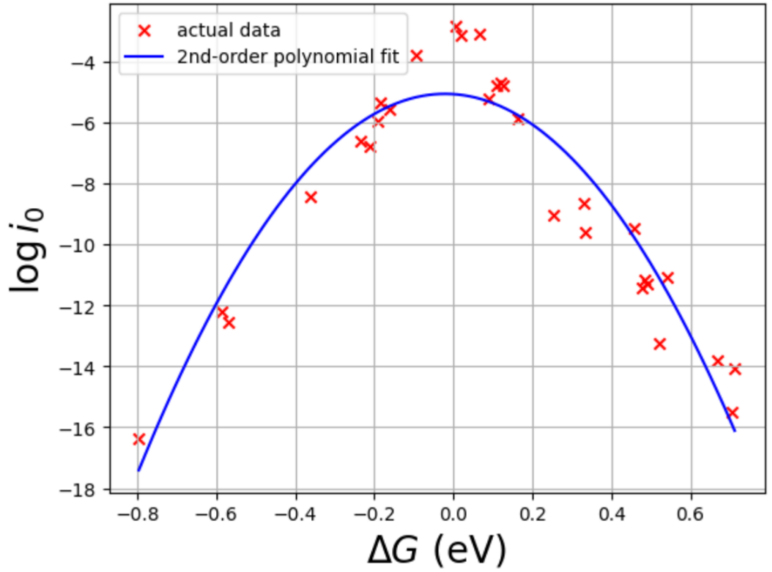


Lesson 9: Optimization

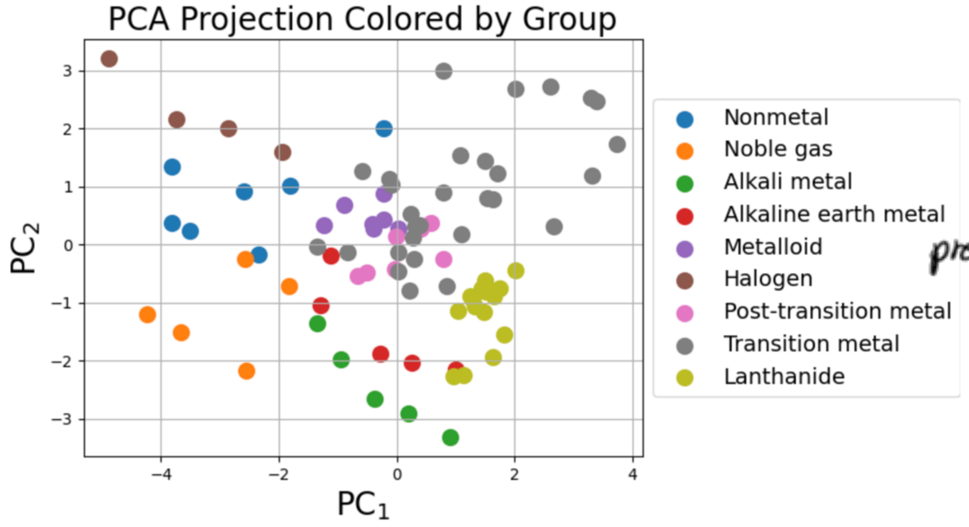
Content: Basic Machine Learning Methods

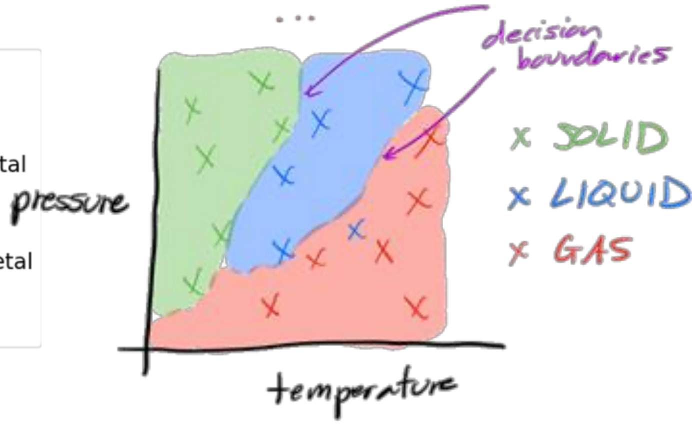
WAYNE STATE UNIVERSITY

Concept	Examples
Regression (linear, multiple, polynomial, ridge, LASSO, logistic)	Control of liquid tank height, Sabatier principle
Principal component analysis	Periodic table analysis
K-nearest neighbors, K-means	Phase classification



Lesson 6: Polynomial Regression

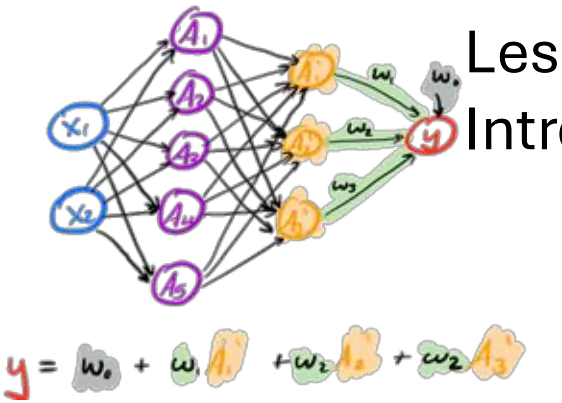




Lesson 20: PCA, KNN, K-Means

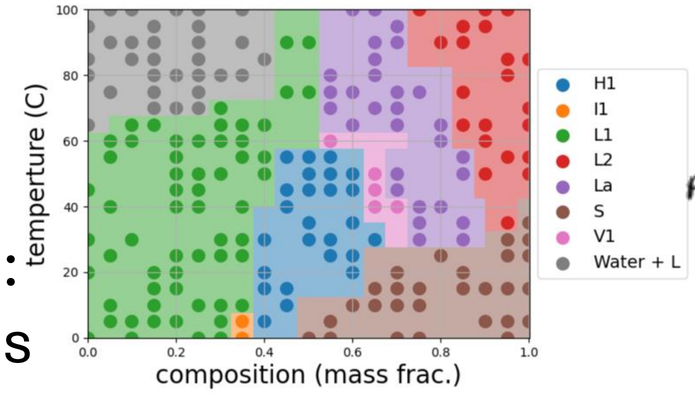
Content: Advanced ML Methods

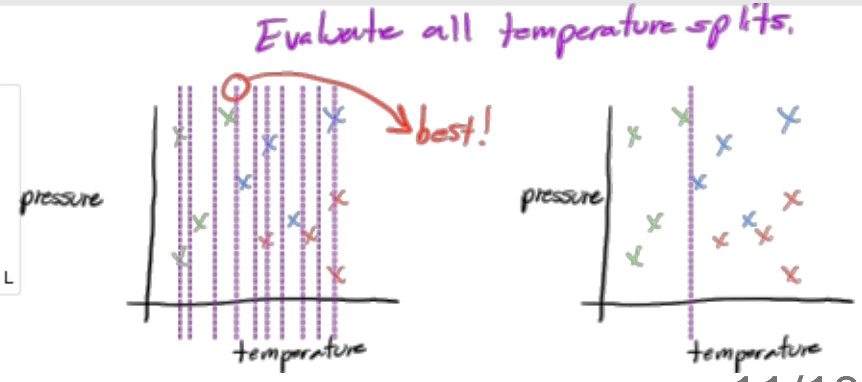
Concept	Examples
Auto-differentiation, backprop.	
Neural networks (dense, convolutional, recurrent)	MOF H_2 loading prediction, small molecule solubility prediction
Discriminant analysis	Periodic table classification
Decision trees, random forests	Surfactant phase behavior
Molecular simulation	Covid-19 protein spike, machine-learned potentials
AlphaFold	Protein structure from sequences



Lesson 10:
Intro to NNs

Lesson 23:
Decision Trees





Cooper & Colón, *MSDE*, 2023; Thacker *et al.*, *JPC B*, 2023

Tip: Run AlphaFold Easily with ColabFold

WAYNE STATE UNIVERSITY

ColabFold v1.5.5: AlphaFold2 using MMseqs2

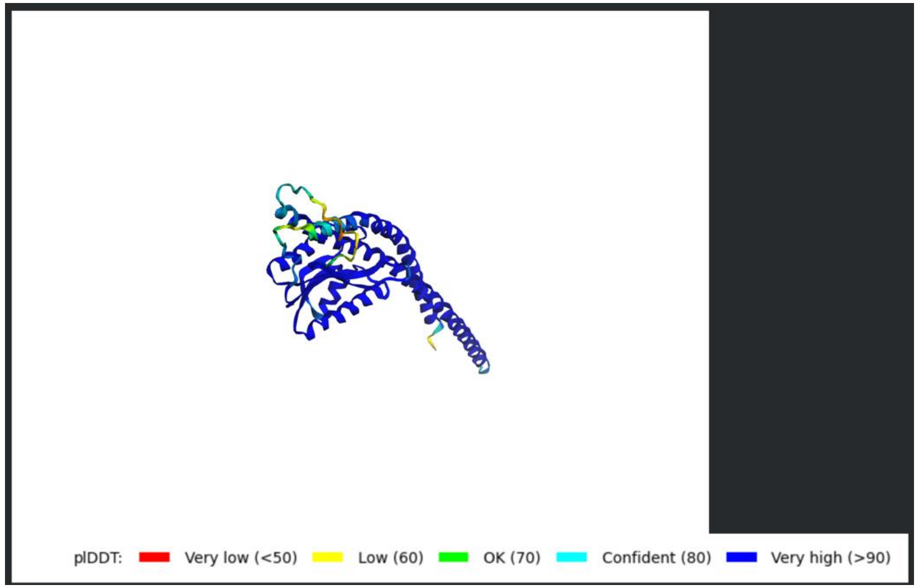
Easy to use protein structure and complex prediction using AlphaFold2 and Alphafold2-multimer. Sequence alignments/templates are generated through MMseqs2 and HHsearch. For more details, see bottom of the notebook, checkout the ColabFold GitHub and Nature Protocols.

Old versions: v1.4, v1.5.1, v1.5.2, v1.5.3-patch

Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. ColabFold: Making protein folding accessible to all. *Nature Methods*, 2022

Input protein sequence(s), then hit `Runtime` -> `Run all`

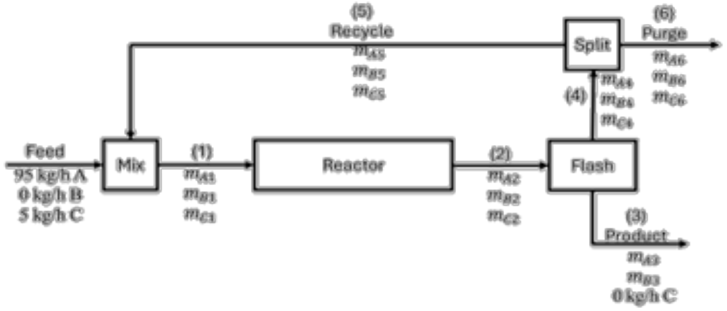
query_sequence: " MAGAKDIRSKIASVQNTQKITKAMEMVAASKMRKSQDRMAASRPYAETMRKVIGHLAHGNLEYKHPYLEDRDVKRVGYLVVSTDRGLAGGLNINLFKKLLAEMKTWTDK "



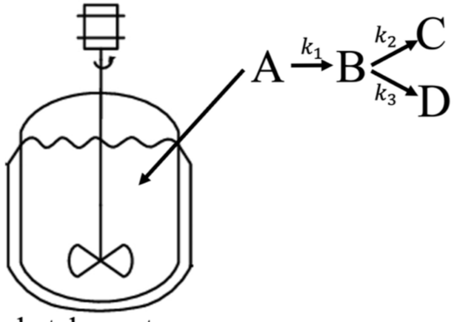
Mirdita, *et al.*, *Nature Methods*, 2022

Relate Assignments to Real Applications

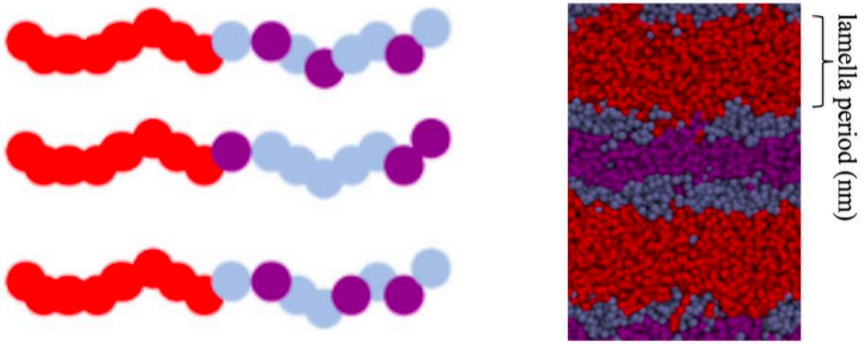
2. **Solving a Process Flow Diagram with Linear Algebra.** A reactor-separator-recycle section is a common paradigm in chemical processing. Calculations involving recycle streams can be a pain to do by hand. In the following process, we feed a stream of mostly A and a little C into a reactor where some A is converted to B. C is inert and does not participate in the reaction. After the reactor, the stream is fed to a flash drum where the bottom stream is a concentrated stream of our product B. The top stream is split into a purge and a recycle. The recycle is mixed with the feed stream and fed back into the reactor. Our mass flow rates are given as m_{Xi} where $X = A, B, C$ is the species and $i = 1, 2, 3, 4, 5, 6$ is the stream number.



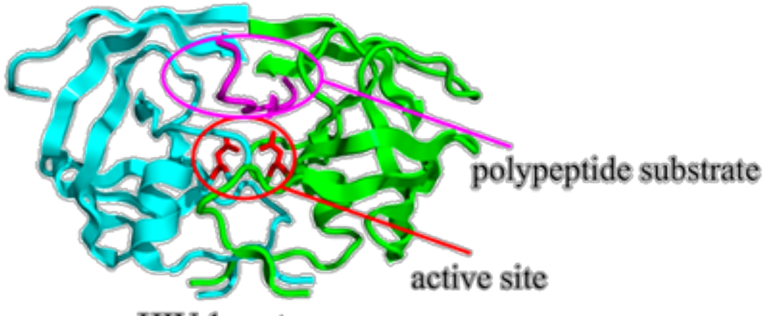
2. **Optimizing a Chemical Reactor.** A batch chemical reactor is a closed reaction vessel that is charged with reactants at time $t = 0$. The reaction proceeds for a certain amount of time before the products are extracted. Using an external jacket, we can control the temperature of the reactor. We are designing a reactor to optimize the depicted reaction scheme. Our reactant is species A, which undergoes a first order reaction with rate constant k_1 into species B. Species B can then undergo a reaction into species C with rate constant k_2 or species D with rate constant k_3. Species B is the desired product and we want to maximize its concentration.



5. *(CHE 7507 only)* **Polymer Layers, Recurrent Layers.** A block polymer consists of a section made of one type of polymer and a section made of a different type of polymer. These block polymers form alternating layers, called lamellae. A computational study looked at how the composition of a block affects the width (period) of these layers based on the composition of one of the block polymer sections. You can find this study here: `https://pubs.acs.org/doi/full/10.1021/acs.macromol.3c02401`



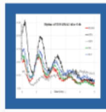
2. **Fighting a Virus with Machine Learning.** Pictured below is HIV-1 protease, a critical component of the human immunodeficiency virus (HIV). HIV-1 protease is an enzyme, which is a protein that catalyzes a chemical reaction. Enzymes have active sites where the chemical reactions take place. Specifically, a protease cuts (cleaves) another protein into pieces. HIV-1 protease is critical in the "life cycle" of the virus because it cuts long proteins into functional pieces. Because it is critical to virus reproduction, HIV-1 protease is a drug target. An inhibitor is a molecule that can bind to an enzyme active site and cause it to stop work, essentially clogging up the enzyme. HIV-1 protease inhibitors are a common class of drug for treating HIV infections.



Mysona, *et al.*, *Macromolecules*, 2024; Rögnvaldsson, *et al.*, *Bioinformatics*, 2015

Tip: UCI ML Repo for Lots of Clean Datasets

UC Irvine
Machine Learning Repository



Toxicity
Donated on 5/4/2022

The dataset includes 171 molecules designed for functional domains of a core clock protein, CRY1, responsible for generating circadian rhythm. 56 of the molecules are toxic and the rest are non-toxic.

Dataset Characteristics	Subject Area	Associated Tasks
Tabular	Biology	Classification

Feature Type	# Instances	# Features
-	171	1203

Superconductivty Data
Donated on 10/11/2018

Two file s contain data on 21263 superconductors and their relevant features.

Dataset Characteristics	Subject Area	Associated Tasks
Multivariate	Physics and Chemistry	Regression

Feature Type	# Instances	# Features
Real	21263	81

HIV-1 protease cleavage
Donated on 4/24/2015

The data contains lists of octamers (8 amino acids) and a flag (-1 or 1) depending on whether HIV-1 protease will cleave in the central position (between amino acids 4 and 5).

Dataset Characteristics	Subject Area	Associated Tasks
Multivariate	Biology	Classification

Feature Type	# Instances	# Features
Categorical	6590	1

archive.ics.uci.edu

AI Use In Class

- LLM use (Claude, Copilot, ChatGPT) is permitted for programming
  - Not a programming course!
  - Introduce students to AI tools
  - Asked students to describe AI use on assignments
  - Use in a controlled environment
- Some problems are marked *“No AI”*
  - Honor system
  - How do I enforce this or design around it?
- Explore how LLMs work
  - Relationship between thermodynamic and LLM temperature
  - Shortcomings and legal issues

Student Feedback

- “I appreciated that we started with learning some basics of python because as a beginner that was interested in the class**, I worried that it would be harder for me to keep up with the coding**.”
- “remove most of the introductory stuff”
- “The scope of practical application, based on this courses learning objectives, has the opportunity to extend far beyond foundational chemical engineering practices. I enjoy this course because it introduces students to coding (Python), machine learning/data analytic techniques, and the **opportunity to experiment with large language models and AI platforms**.”
- “It is really interesting to see the applications of machine learning in a variety of topics, in addition I got to see the relevant topics **combining information from previous classes** such as Schrodinger's equation of states from Physical Chemistry or reaction kinetics and reactor design.”
- “The assignments helped us **experiment and practice** the concepts we learned.”
- “I enjoyed the way your slides on each topic sort of **introduced topics from the ground floor–up**.”

Acknowledgements

- Jeffrey Potoff, Wayne State
- Camille Bishop, Wayne State
- Ali Seyedi, Wayne State
- Yamil Colón, Notre Dame
- Jindal Shah, Oklahoma State
- Bingqing Cheng, UC Berkeley
- Cory Simon, Oregon State
- Teresa Head-Gordon, UC Berkeley

WAYNE STATE UNIVERSITY

AI Tools Are Not Good or Bad, They Are What We Make Them

WAYNE STATE UNIVERSITY

“Commercial Chemical Operations”

“Manufacture of Poison Gas Bombs”

Questions?

Detroit Industry Murals, 1932-1933, Diego Rivera, Detroit Institute of Arts