Introduction to Machine Learning in the Physical Sciences: A New Hands-on Course at Columbia

Simon J. L. Billinge[1], Sanat Kumar[2], Tian Zheng[3], Ruining Zhang[1], Tejus Shastry[2]
[1]Applied Physics and Applied Mathematics, [2]Chemical Engineering, [3]Statistics

Overview and Motivation

What: A one-semester accelerated machine learning (ML) course applied to STEM research problems.

Target Audience: graduate students and senior physical science undergraduates interested in applying ML to their research.

Motivation: create a hands-on course that teaches the application of ML to physical science research problems; prepare students for the increasing impact of AI and ML on physical sciences.

Scope: unsupervised and supervised learning, decision trees, logistic regression, neural nets, etc., applied to physical science research problems

Prerequisites: Physical science knowledge and basic programming. Prior ML knowledge is NOT required.

Hands-on educational examples: EdEx

What: EdEx stand for educational examples. They are hands-on tutorials designed to walk students through machine-learning applications from published research.

EdExes:

- integrate with the theoretical lecture component of the course
- are based on published research conducted primarily here at Columbia Engineering
- Contain two jupyter notebooks: one is the "solution," where the problem is fully presented and solved along with detailed explanations; the other is the "problem set" where the students are guided to fill in the incomplete code

Tools introduced: Jupyter notebooks, conda, scikit-learn, keras, tensorFlow.

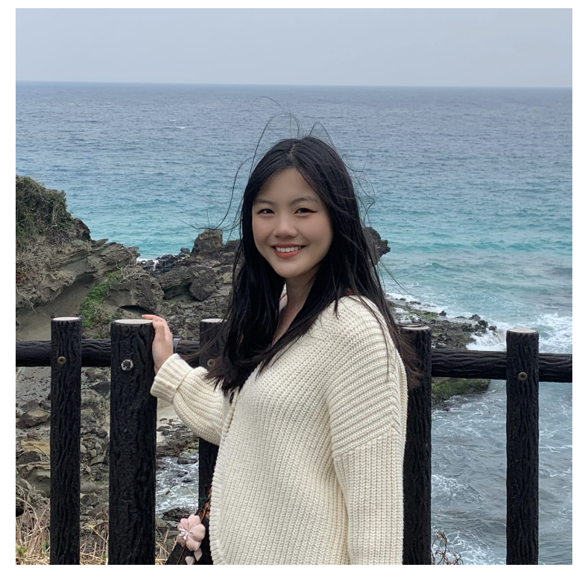
Status and Team

Status: Prototype EdExs will be given in Summer A 2023. First full course expected to be run in spring 2024

Funded by:

COLLABORATORY AT COLUMBIA UNIVERSITY

Ruining Zhang (MatSci PhD)



Tejus Shastry (ChemE PhD)

EdEx 1: Space Group Classifier

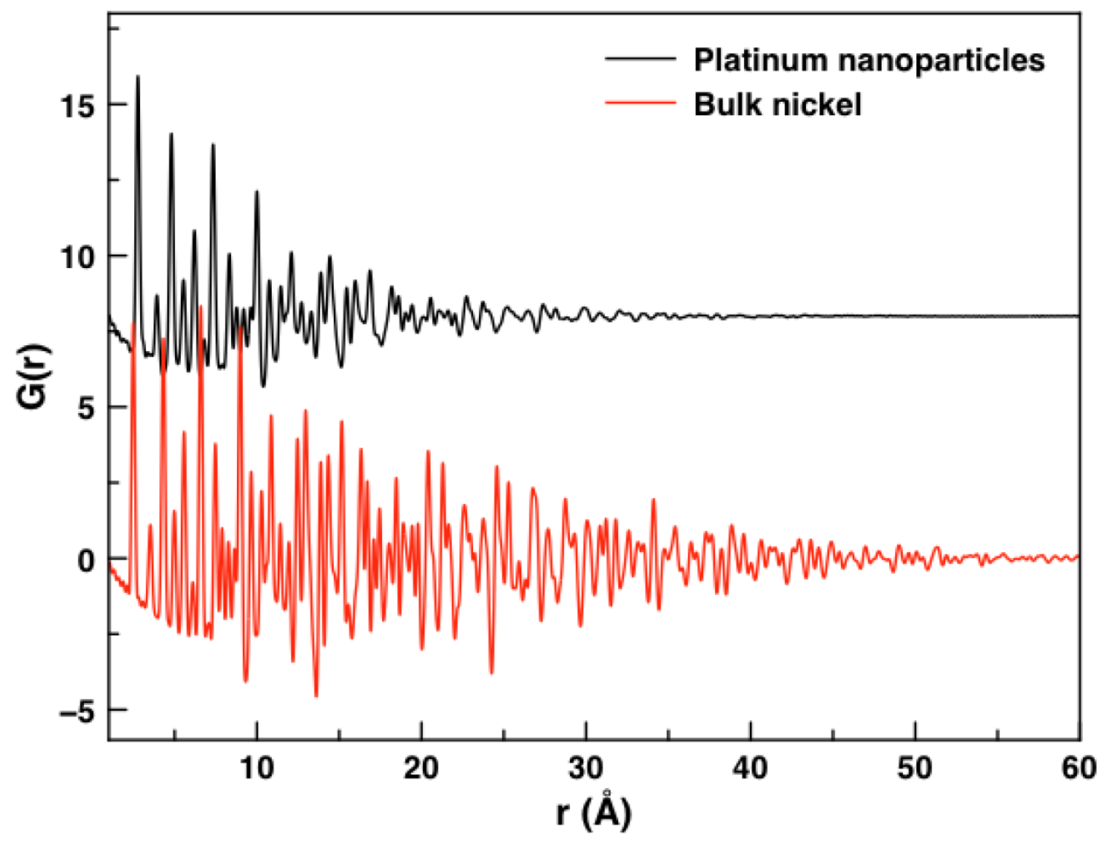


Image source: Billinge Group

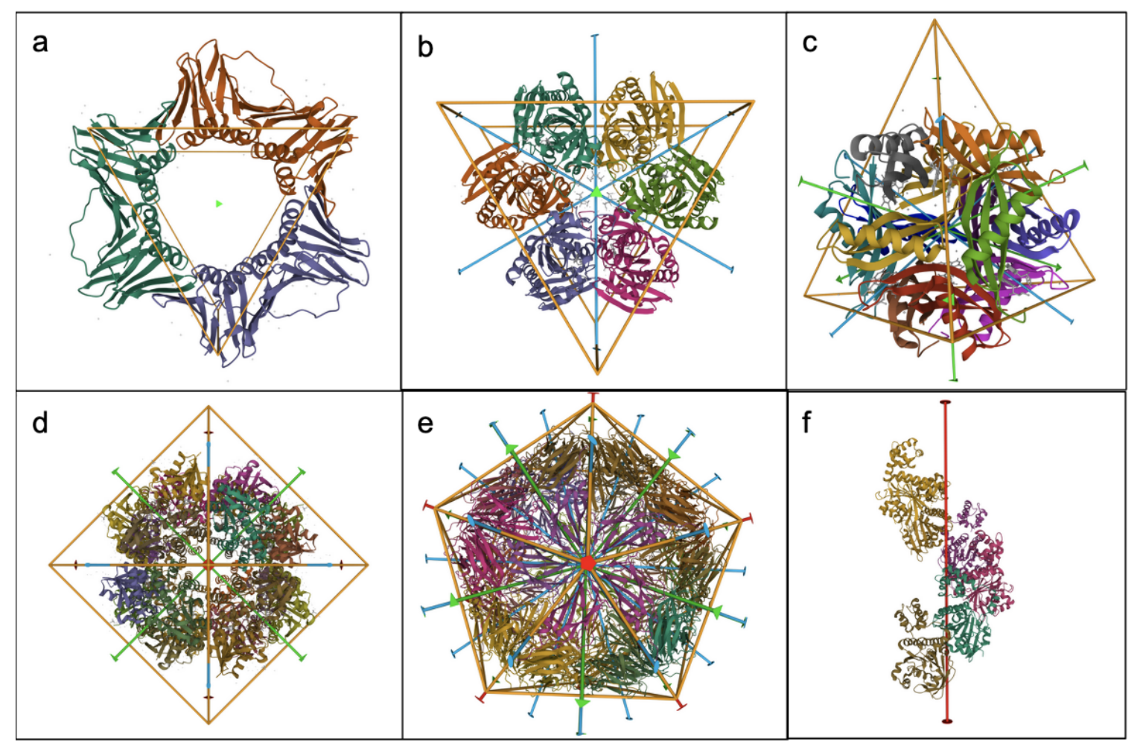


Image source: RCSB Protein Databank

Goal: predict the space group (SG) of a crystal structure giving the measured atomic pair distribution function (PDF).

Problem: SG encodes structural symmetries of atomic arrangements in a material. PDF is an x-ray measurement of the material. There is no direct way of getting SG from the PDF.

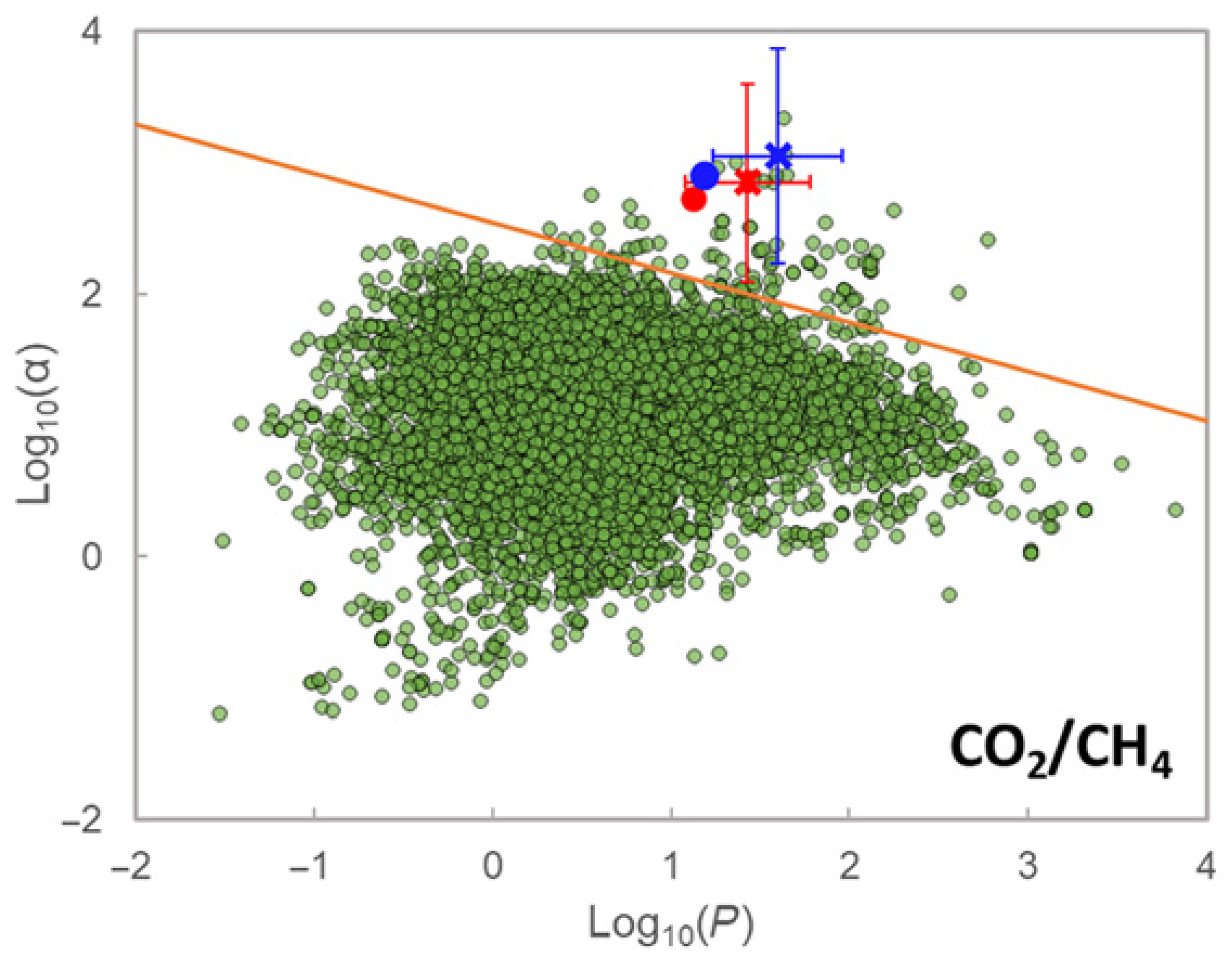
Motivation: The ML model can quickly predict the most likely space groups and give insights into the structure-property relationships.

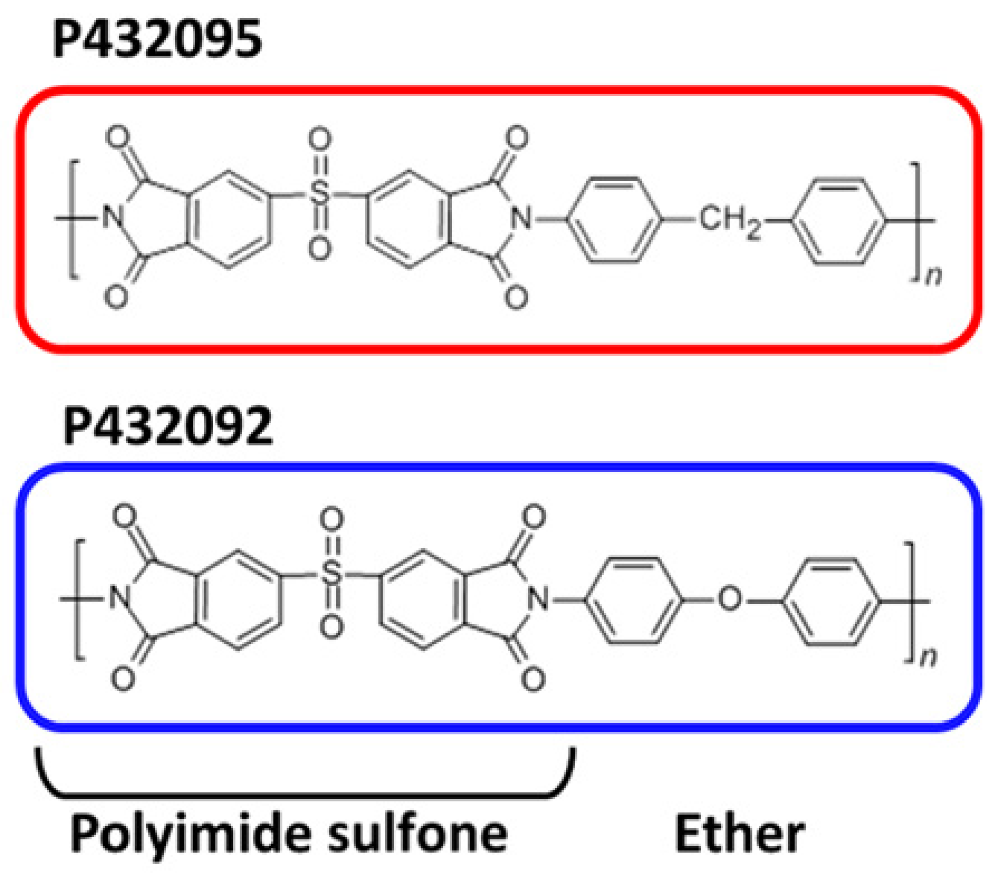
Model: convolutional neural networks.

Training: 40,000 PDFs that are calculated from 8 of the most common space groups.

Liu, C. H., et al. Acta Crystallographica Section A: Foundations and Advances, 75(4), 633-643.

EdEx 2,3: Membrane Permeability





Goal: predict the gas phase separation properties of polymers for 6 gases from only monomer chemistry; understand which monomer chemistries perform well.

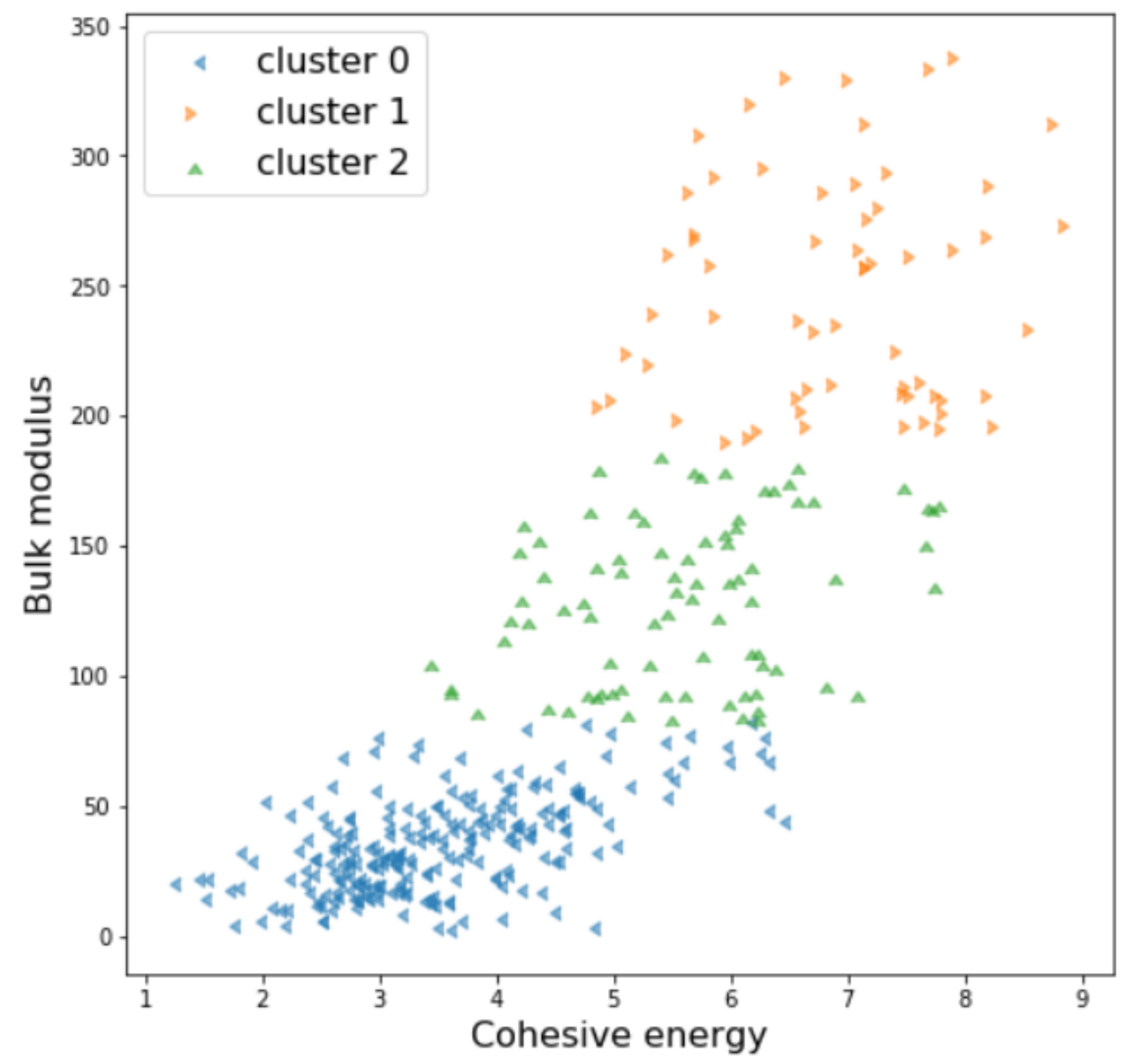
Problem: We need to separate individual gases from mixtures at scale, e.g. CO_2 from the air. We need more selective and permeable organic polymers to do it.

Motivation: can ML predict which known polymers will perform well without actual testing? Probe the underlying physics of gas transport and identify potential top-performing chemical motifs for next-generation separations research?

Model: various regression models implemented via scikit-learn; permutation importance, Shapley Values.

Barnett, J. Wesley, et al. Science Advances, vol. 6, no. 20, 2020.

EdEx 4: Predicting melting temperatur



Goal: Predict the melting temperature of inorganic materials given just the chemistry of the constituents.

Problem: We need quick and reliable low-cost predictions of melting temperature for metal extraction. Chemical variability means that one model does not work over high-variance data.

Motivation: Can we use clustering to pre-sort data so different models can be used for different chemistries?

Model: k-Means clustering, various other clustering and regression models implemented via scikit-learn

Gharakhanyan, V., Urban, A., In preparation (2023).

Columbia University
Data Science Institute