AI-Driven Prediction of Material Properties

Description

In this scientific project, your task is to build a machine learning model for predicting various properties of materials. This project is closely related to key questions in materials science, which revolves around understanding, designing, and manipulating materials to achieve desired properties and functionalities. In particular, you will address the structure–property relationship: how does the structure of a material influence its macroscopic properties (mechanical, electrical, thermal, optical)?

Over the course of this project, you will put into practice the machine learning methods introduced in the lecture. You will learn how to analyze data on materials and build a machine learning model from scratch to predict material properties.

As a prerequisite, you should have basic knowledge of programming in Python. For building machine learning models, we will use the PyTorch library (https://pytorch.org/). It is recommended to do the coding work in a Jupyter Notebook (https://jupyter.org/). The dataset we will use is the *matminer* (https://matminer.readthedocs.io/en/latest/) collection of materials properties.

Your overall task is to build machine learning models to predict the properties of a class of materials based on specific features. The individual projects described below differ only in terms of the specific dataset. Depending on the particular project you choose from the list below, the features will be either the composition or the structure of the material. Your task is divided into three main subtasks:

1. Construct a **linear regression model** for predicting material properties.
2. Construct a **neural network model** for predicting material properties.
3. Assess the accuracy of both models.

The following steps will guide you through the process of building your machine learning models.

Data preparation

Analyze the raw dataset by inspecting the distribution of features and target properties — these are the inputs and outputs of your machine learning model. Visualize these relationships to gain a better understanding of your data. Use the tools provided in *matminer* to featurize your data and visualize the featurized dataset. Consider rescaling your data if needed.

Model choice

As discussed in the lecture, the modeling task falls under the category of supervised machine learning. You will construct two models: a linear regression model and a neural network model. You can use PyTorch to build both models.

Linear regression model

Linear regression is a fundamental statistical technique used to model and analyze the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data. The simplest form, called simple linear regression, involves a single predictor and estimates the best-fit line using the least-squares method, minimizing the sum of squared differences between observed and predicted values.

The linear regression model will serve as a baseline — you will compare your neural network model against its results.

Neural network model

Neural network model training and optimization

Split your dataset into training, validation, and test sets. Before doing so, reflect on why data splitting is necessary: training and testing a machine learning model on the same data would be a methodological error. When splitting the dataset, consider appropriate ratios for training, validation, and test sets. Also note: the dataset will determine the input and output dimensions of your neural network.

Train your neural network model and monitor the training and validation losses. What do these losses indicate? Make sure your model is neither underfitting nor overfitting.

Overfitting is common and can be addressed through various strategies, including regularization (L1, L2, dropout), early stopping, batch normalization, and k-fold cross-validation. Familiarize yourself with at least two of these methods, retrain your model using them, and compare the training progress of your models.

In addition to the model parameters optimized during training via backpropagation, neural networks also require optimization of hyperparameters, which are tuned outside the training loop. Examples include the network architecture (number of hidden layers and neurons), learning rate, batch size, activation function, and optimizer choice. Conduct a hyperparameter study by varying at least three of these hyperparameters. Save and compare your resulting neural network models.

By following the steps above, you should have trained several neural networks. Save at least five models, each differing in hyperparameter settings.

Neural network model testing

Now identify your best-performing model by evaluating all saved models on the unseen test set. Assess their accuracy and determine the best one.

How can you evaluate the stability of your model's predictions? (Hint: recall that neural networks are initialized with random weights — you can use this to test prediction stability.)

Model assessment

Compare the prediction accuracy of your best neural network model with that of the linear regression model.

Scientific Project 11: Band Gap of Perovskites

Perovskites are a class of materials that share a specific crystal structure, typically represented by the formula ABX_3, where 'A' and 'B' are cations of different sizes and 'X' is an anion, often oxygen or a halide. This versatile structure allows for a wide range of chemical compositions and tunable properties. In particular, metal halide perovskites have gained significant attention for their exceptional light absorption, charge transport, and low-cost synthesis, positioning them as promising materials for high-efficiency solar cells. In the context of perovskites, the band gap refers to the energy difference between the material's valence band (occupied by electrons) and its conduction band (where electrons can move freely and conduct electricity). This property determines how a perovskite interacts with light and electricity, making it crucial for applications in solar cells, LEDs, and other optoelectronic devices. A suitable band gap allows perovskites to efficiently absorb sunlight and convert it into electricity, which is why tuning the band gap is a key focus in perovskite materials research for photovoltaics.

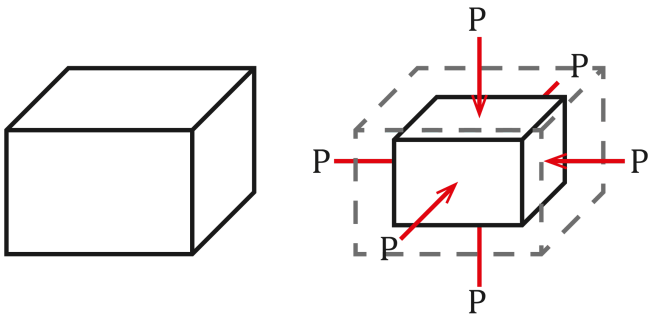
Using the *castelli_perovskites* dataset in *matminer*, build a machine learning model to predict the band gap of perovskites based on their composition, structure, or both.

Perovskite [Link to image].

Scientific Project 12: Bulk Modulus

The bulk modulus is a measure of a material's resistance to uniform compression. It quantifies how much a material will compress under pressure and is defined as the ratio of the applied pressure to the resulting relative decrease in volume. A high bulk modulus indicates that a material is incompressible and stiff, while a low bulk modulus means the material is more compressible. This property is important in materials science and engineering, especially for evaluating the mechanical stability of solids under various conditions such as pressure or structural load.

Using the *matbench_log_kvrh* dataset in *matminer*, build a machine learning model to predict the bulk modulus of materials based on their structure.

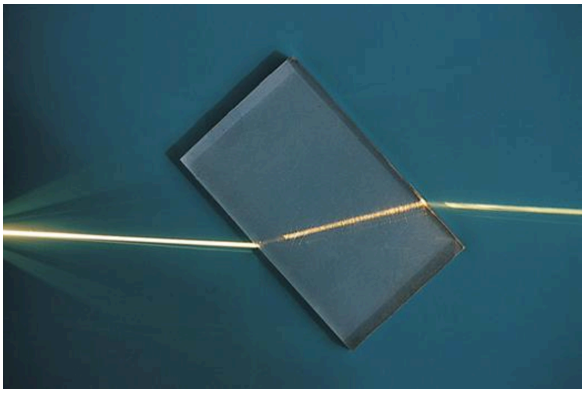


Deformation of a parallelepiped through isostatic pressure [Link to image].

Scientific Project 13: Refractive Index

The refractive index is a measure of how much light bends, or refracts, as it passes from one medium into another. It is defined as the ratio of the speed of light in vacuum to its speed in the material. A higher refractive index means that light travels more slowly through the material and bends more sharply at the interface. This property is crucial in optics and photonics, as it determines how materials interact with light.

Using the *matbench_dielectric* dataset in *matminer*, build a machine learning model to predict the refractive index of materials based on their structure.

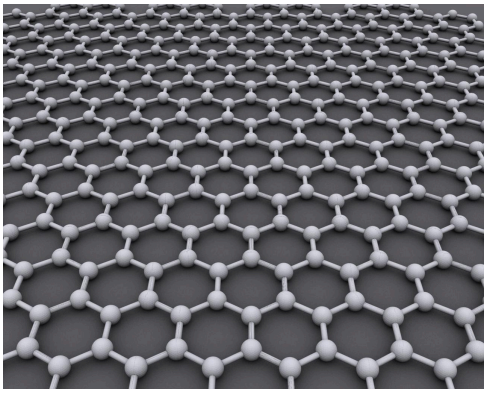


A ray of light being refracted in a plastic block [Link to image].

Scientific Project 14: Exfoliation Energy of 2D Materials

Exfoliation energy refers to the amount of energy required to separate a single atomic layer from a bulk layered material to obtain a two-dimensional (2D) sheet. It is a key parameter in the study and fabrication of 2D materials, such as graphene or transition metal dichalcogenides, which are known for their unique electronic, mechanical, and optical properties. A lower exfoliation energy indicates that a material can be more easily peeled into stable monolayers, making it more suitable for applications in nanoelectronics, flexible devices, and energy storage technologies.

Using the *jarvis_ml_dft_training* dataset in *matminer*, build a machine learning model to predict the exfoliation energy of 2D materials based on their composition, structure, or both.

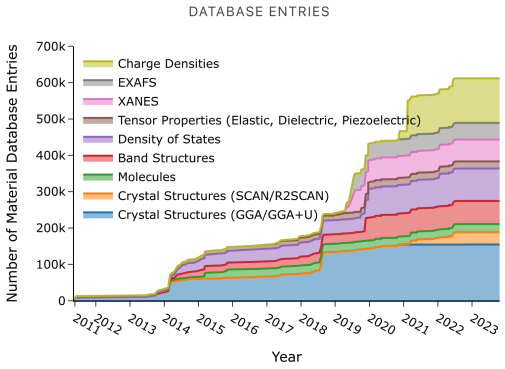


The ideal crystalline structure of graphene in a hexagonal grid [Link to image].

Scientific Project 15: Formation Energy

Formation energy is the energy change associated with forming a compound from its constituent elements. It provides a measure of the thermodynamic stability of a material: a negative formation energy indicates that the compound is energetically favorable and likely to form spontaneously, while a positive value suggests instability. In materials science, formation energy is used to predict which compounds are stable, compare the stability of different phases, and guide the discovery of new materials.

Using the *matbench_mp_e_form* dataset in *matminer*, build a machine learning model to predict the formation energy of materials based on their structure.

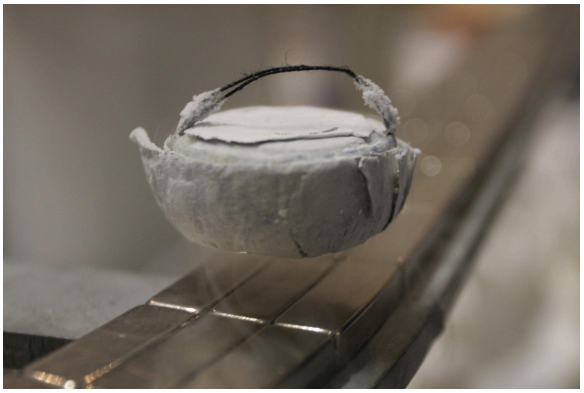


Materials Project [Link to image].

Scientific Project 16: Superconductivity

A superconducting material can conduct electric current with zero electrical resistance when cooled below a certain critical temperature. This phenomenon allows superconductors to carry electricity without energy loss, making them highly efficient for applications such as power transmission, magnetic levitation, and advanced medical imaging systems like MRI. Research into superconducting materials focuses on discovering compounds that exhibit this behavior at higher temperatures, making them more practical for widespread use.

Using the *superconductivity2018* dataset in *matminer*, build a machine learning model to predict the superconducting temperature of materials based on their composition.



A high-temperature superconductor levitating above a magnet [Link to image].