Prediction of Inorganic Crystals' Thermodynamic Stability Using Machine Learning

Sara Nabili

July 11, 2025

Outline

- Intro
- Methodology
- Feature
- Introduction
- Feature Selection
- TODOs
- Methodology
- ML algorithms
- DNN
  - DNN - Assessment
  - DNN - Partial dependence
    - DNN - PDP plots
- results and conclusion
- References
- Backup Slides

Introduction

Purpose: This project uses machine learning (ML) binary classification algorithm to predict the thermodynamic stability of inorganic crystals based on various features.

Motivation: In materials science, predicting compound stability is costly and slow. This project aims to accelerate discovery using data-driven models.

Tools/Frameworks: Python, scikit-learn, TensorFlow, XGBoost, pymatgen, data-analysis toolkits, visualization packages, parallel and multiprocessing computing frameworks

Implementation: Feature extracted using pymatgen, engineered descriptors like bond statistics and atomic fractions, and trained ML algorithms such as a deep neural network.

Final Result: The final model achieved 75% validation accuracy, showing promising results for early-stage stability screening.
Application: This tool can be used for material engineering, pharmaceutical, and sustainable energy technologies

General Information

Problem Context: Thermodynamic stability (under a set of conditions) is achieved if the formation energy of the compound can not be lowered by rearranging its atoms [1].

Energy lowering mechanism:

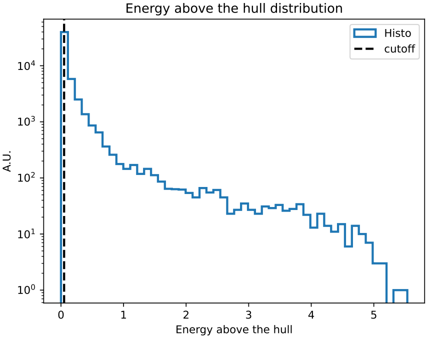
- *Decomposition:* Phase separation into competing materials that share an identical overall composition
- *Polymorphism:* Phase transition into an alternative crystal structure (polymorph) at fixed composition.

Thermodynamic Stability:

- A convex hull represents the lowest formation energy envelope derived from inorganic crystalline materials plotted against their composition.
- *The energy above the hull (EBH):* The energy difference between a compound and the convex hull with the same composition.

Energy above the hull

- The ML algorithms are trained to predict the thermodynamic stability of chemical compounds by the binary classification of the energy above the hull feature (EBH) [1]. The EBH distribution is shown below.
- In this study, thermodynamically stable to slightly metastable materials are classified by applying a threshold of $\mathrm{EBH} \leq 0.05$ as suggested by [4].
- Out of nearly 50,000 materials in the dataset, roughly half meet the criteria for thermodynamic stability, while the other half fall above the defined threshold and are considered unstable.



Limitations of DFT and the Rationale for Machine Learning

Limitations of Density Functional Theory (DFT):

- Highly accurate, but computationally intensive, especially for large or low symmetry structures.
- High-throughput screening of new compounds is limited by cost and runtime.
- Not suited for rapid prototyping or exploration of vast chemical spaces.

How Machine Learning Addresses These Bottlenecks:

- Learns structure property relationships from curated datasets.
- Predicts material stability orders of magnitude faster than DFT
- Enables scalable screening of millions of candidate materials

Study Limitations and Future Directions

Limitations:

- **Dataset Scale:** Access to the Materials Project database was limited to 50,000 compounds, constraining the deep neural network's generalization capacity.
- **Computational Resources:** Experiments were conducted on a single machine with 8 CPU cores, limiting parallelization and throughput.

Future Directions:

- Expanding access to larger datasets can significantly enhance model performance and diversity.
- Deploying high-performance clusters and distributed computing frameworks (e.g., HTCondor) will support advanced feature engineering and more complex ML architectures.

Data Collection

Data source: Data extracted from the Materials Project [1], an open database of DFT-computed material properties.

Feature extraction: The features are classified into four classes:

- **Energy & electronic features:** Helps capture energetic contribution
- **Structural and composition:** Gives access to size, packaging, and complexity
- **Bond features:** Facilitates the detection and quantification of inter-element bonding, while representing the surrounding chemical context
- **Atomic fraction:** Reflects elemental influence on stability

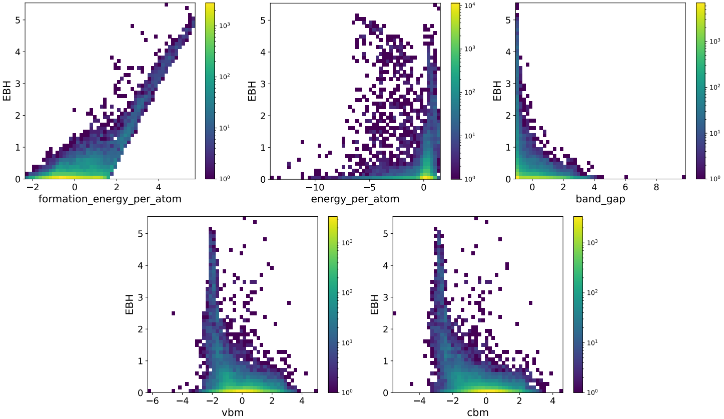
The first two classes are database queries, whereas the rest need feature engineering.

Feature Engineering To build:

- Atomic fractions: Computed element-wise from compositions.
- Bond structure statistics: Using neighborhood-finding algorithms, MinimumDistanceNN [8]:
  - Number of bonds
  - Mean bond length
  - Bond length standard deviation

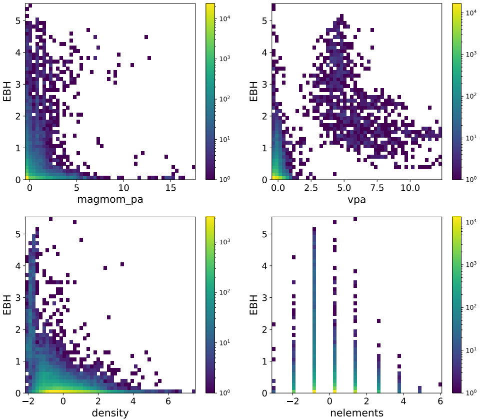
Feature: Energy & Electronic Features

- Formation energy per atom: core feature
- Energy per atom: include element reference in formation energy per atom
- Band gap: related to chemical bonding and electronic stability
- Valence/conduction band edges: explores patterns in electronic structure



Feature: Structure and Composition

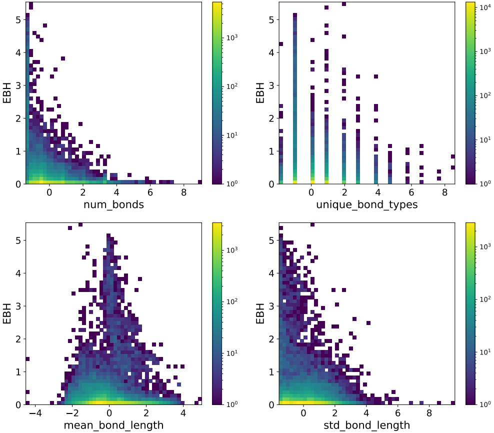
- Magnetic moment per atom: Depending on the material, some stable phases are magnetic
- Volume per atom: global structure feature
- Density: affected by atomic mass and volume
- Number of elements and sites: Capture compound's complexity; too many elements and sites may lead to less stable/metastable phases. Number of sites is considered in magnetic moment and volume per atom computation (see backup slide).



Feature: Bond structure

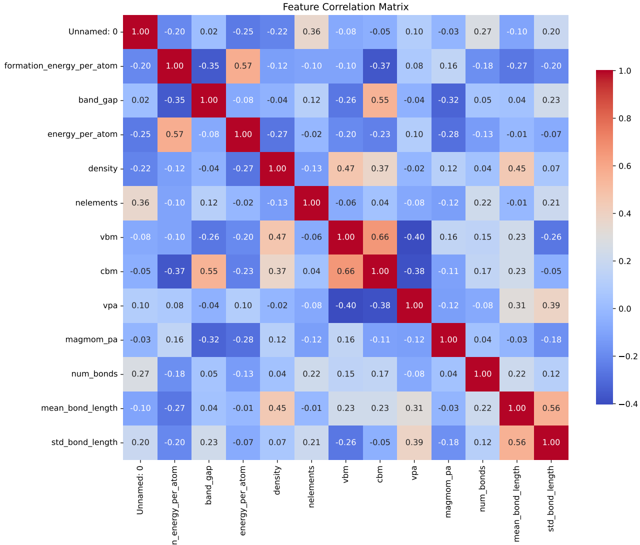
Captures bond's (min/avg/max) distance and coordination using MinimumDistanceNN [8]. Features are:

- Number of bonds: coordination and connectivity
- Unique bond type: structural diversity and bonding motifs
- Bond length mean and standard deviation: average bonding length and geometrical distortion



Correlation Matrix

XGBoost [2] was applied to analyze relationships among the ML input features. Their pairwise correlations are visualized as a correlation matrix.



	Unnamed: 0	n_energy_per_atom	band_gap	energy_per_atom	density	nelements	vbm	cbm	vpa	magmom_pa	num_bonds	mean_bond_length	std_bond_length
Unnamed: 0	1.00	-0.20	0.02	-0.25	-0.22	0.36	-0.08	-0.05	0.10	-0.03	0.27	-0.10	0.20
formation_energy_per_atom	-0.20	1.00	-0.35	0.57	-0.12	-0.10	-0.10	-0.37	0.08	0.16	-0.18	-0.27	-0.20
band_gap	0.02	-0.35	1.00	-0.08	-0.04	0.12	-0.26	0.55	-0.04	-0.32	0.05	0.04	0.23
energy_per_atom	-0.25	0.57	-0.08	1.00	-0.27	-0.02	-0.20	-0.23	0.10	-0.28	-0.13	-0.01	-0.07
density	-0.22	-0.12	-0.04	-0.27	1.00	-0.13	0.47	0.37	-0.02	0.12	0.04	0.45	0.07
nelements	0.36	-0.10	0.12	-0.02	-0.13	1.00	-0.06	0.04	-0.08	-0.12	0.22	-0.01	0.21
vbm	-0.08	-0.10	-0.26	-0.20	0.47	-0.06	1.00	0.66	-0.40	0.16	0.15	0.23	-0.26
cbm	-0.05	-0.37	0.55	-0.23	0.37	0.04	0.66	1.00	-0.38	-0.11	0.17	0.23	-0.05
vpa	0.10	0.08	-0.04	0.10	-0.02	-0.08	-0.40	-0.38	1.00	-0.12	-0.08	0.31	0.39
magmom_pa	-0.03	0.16	-0.32	-0.28	0.12	-0.12	0.16	-0.11	-0.12	1.00	0.04	-0.03	-0.18
num_bonds	0.27	-0.18	0.05	-0.13	0.04	0.22	0.15	0.17	-0.08	0.04	1.00	0.22	0.12
mean_bond_length	-0.10	-0.27	0.04	-0.01	0.45	-0.01	0.23	0.23	0.31	-0.03	0.22	1.00	0.56
std_bond_length	0.20	-0.20	0.23	-0.07	0.07	0.21	-0.26	-0.05	0.39	-0.18	0.12	0.56	1.00

Feature Importance

Blank

Blank

Blank

Introduction

Purpose: Train Machine Learning (ML) algorithms to predict the chemical compounds' thermodynamic stability, following the idea of [3] using material project repository [7] MPR database.

Used methods and algorithms: Binary classification to train, validate, and test ML algorithms:

- Logistic Regression
- Random Forest
- Deep Neural Network (DNN)

Application: This tool can be used for material engineering, pharmaceutical, and sustainable energy technologies

General Info

- ML algorithms are used to train, evaluate, and test three ML algorithms to assess chemical compounds' stability prediction, with the focus on DNN. The binary classification is made using the computed energy_above_the_hull from the MPR dataset.
- Dataset: MPR dataset (computationally derived) is used via a free API_KEY.
- All the codes to make this study are made public and pushed to [6] GitHub.

Methodology

- Publicly available dataset from Material Project (MPR) Database [5]: uses computational material science knowledge and computer science techniques to compute properties of materials
- Data Mining:
  - Feature selection is defined in more detail in feature selection slide
  - Filtering datasets: number of sites and volume cuts
  - Computed arithmetically: volume and magnetic moment per number of sites of the crystal:

$$\text{vpa} = \frac{\text{volume}}{\text{nsites}}$$

$$\text{magmom} = \frac{\text{total magnetization}}{\text{nsites}}$$

filtering datasets

- feature selections

Challenges and Limitations

- Dataset: MPR dataset (computationally derived) has a limited number of chemical compounds, thus understanding chemical information and creativity in designing neural net are crucial to get a suitable performance
- Inorganic crystal structural database this is not a limitation –¿ should say that in general information
- Computer power: Only one computer was used to process the data and to optimize hyperparameters of the neural network. should be more specific about computer power: how many CPU and GPU were used, how much CPU hour did i use to train and to compute not existed information
- Lack of some MPR data due to incompatibility between computational and experimental data. This was observed for oxidation state for a number of compounds.

Feature Selection

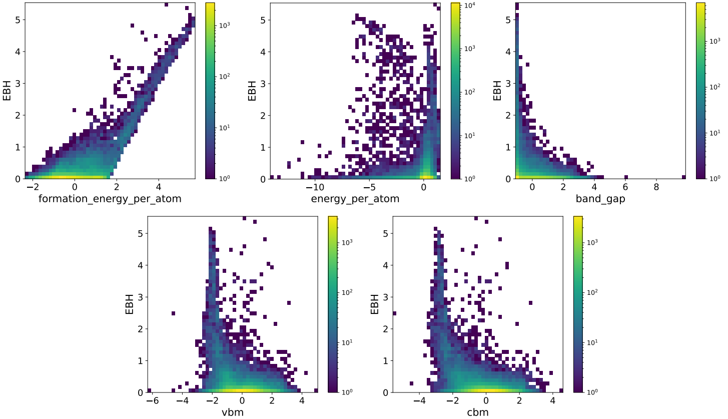
To predict thermodynamic stability, these four feature classes are selected:

- Energy and electronic features: to capture energetic contribution
- Bond features: to capture the presence and proportion of element-wise bonds and to encode local chemical environment.
- Structural and Composition Features: to give access to size, packaging, and complexity
- Elemental statistics:

Feature Selection - Energy & Electronic Features

Helping the model to infer its locality in the energy convex hall.

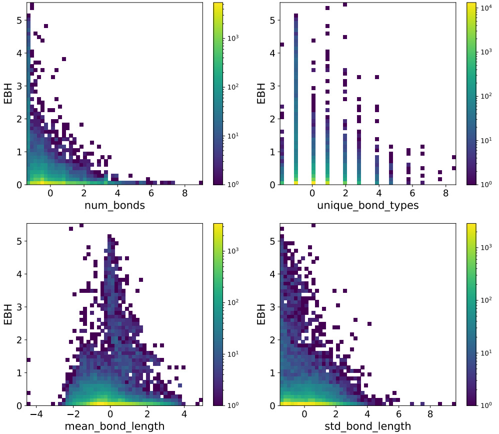
- Formation energy per atom: core feature
- Energy per atom: include element reference in formation energy per atom
- Band gap: related to chemical bonding and electronic stability
- Valence/conduction band edges: explores patterns in electronic structure



Feature Selection - Bond structure

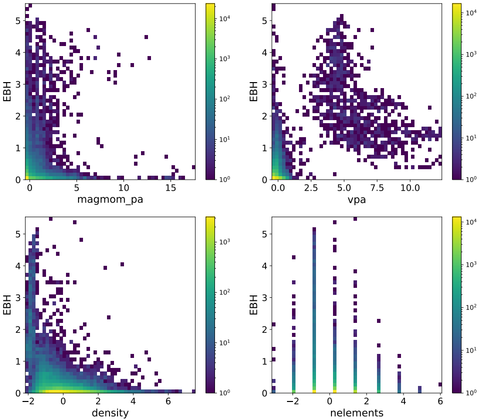
Captures bond's (min/avg/max) distance and coordination using MinimumDistanceNN [8]. Features are:

- Number of bonds: coordination and connectivity
- Unique bond type: structural diversity and bonding motifs
- Bond length mean and standard deviation: average bonding length and geometrical distortion



Feature Selection: Structural and Composition

- Magnetic moment per atom: Depending on the material, some stable phases are magnetic
- Volume per atom: global structure feature
- Density: affected by atomic mass and volume
- Number of elements and sites: Capture compound's complexity; too many elements and sites may lead to less stable/metastable phases. Number of sites is considered in magnetic moment and volume per atom computation (see backup slide).



what to add more

TODO: add how I decided to select features for DNN textcolorredexplain what is energy above the hull; why it affects stability TODO: add the DNN structure used in this study, aka number of neurons, number of layers, number of epochs, loss function

Methodology

- What features are selected from datasets
- how other absent features are computed from the ones present
- How are bond structure and atomic fraction computed from information in MPR
- what type of filters are applied
- selected features that have the least correlation
- features passing feature importance are selected

Logistic Regression

- used logistic regression as a fast to train and interpretable model to check the results from the DNN model for the binary classification as a baseline model.
- The ML algorithms are trained to predict the thermodynamic stability by classifying the energy_above_hull feature. The energy_above_hull distribution is shown below.
- A cutoff of 0.05 eV/atom on energy_above_hull is used, as suggested by [4].

Random Forest

- The ML algorithms are trained to predict the thermodynamic stability by classifying the energy_above_hull feature. The energy_above_hull distribution is shown below.
- A cutoff of 0.05 eV/atom on energy_above_hull is used, as suggested by [4].

XGBoost

- The ML algorithms are trained to predict the thermodynamic stability by classifying the energy_above_hull feature. The energy_above_hull distribution is shown below.
- A cutoff of 0.05 eV/atom on energy_above_hull is used, as suggested by [4].

Deep Neural Net - Parameters

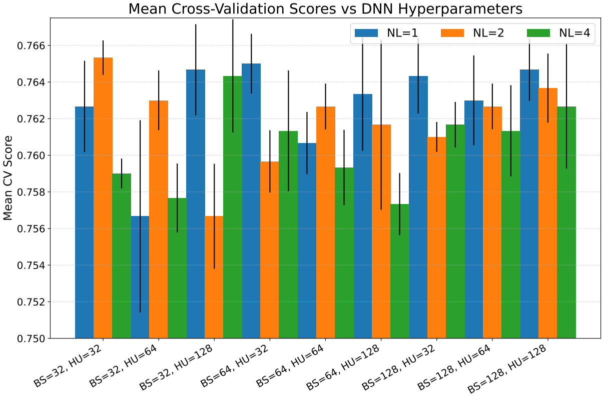
- Tensorflow is used to build the neural network
- Features are selected after their importance passes a certain threshold
- Used Rectified Linear Unit (ReLU) activation function (ReLu(x) = max(0,x)) for the input and hidden layers, and sigmoid activation function for the output layer.
- Prevent overfitting during training:
  - Used random dropout of neurons for the input (10%) and hidden layers (30%);
  - early stopping to prevent validation loss increment;
  - batch normalization to stabilize learning by normalizing input per layer to ensure that the inputs have zero mean and unit variance;
- Learning rate to optimize model weights in response to the estimated error (loss)

Deep Neural Net - Hyperparameter Optimization

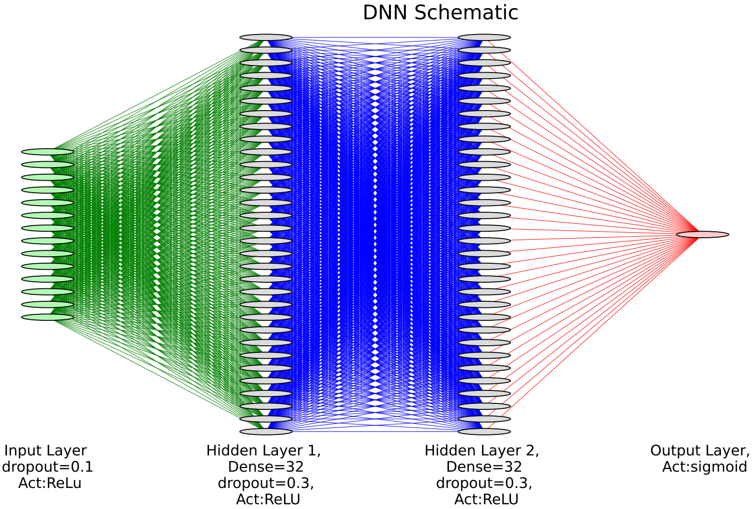
DNN parameters are optimized using a grid search method on the following parameters:

- Number of hidden layers
- Number of neurons per layer
- Batch size

The mean of the three cross-validation scores for the above alternati



Deep Neural Net - Architecture



Deep Neural Net - Performance

- ROC curves
- accuracy
- loss functions

Deep Neural Net - Partial Dependence

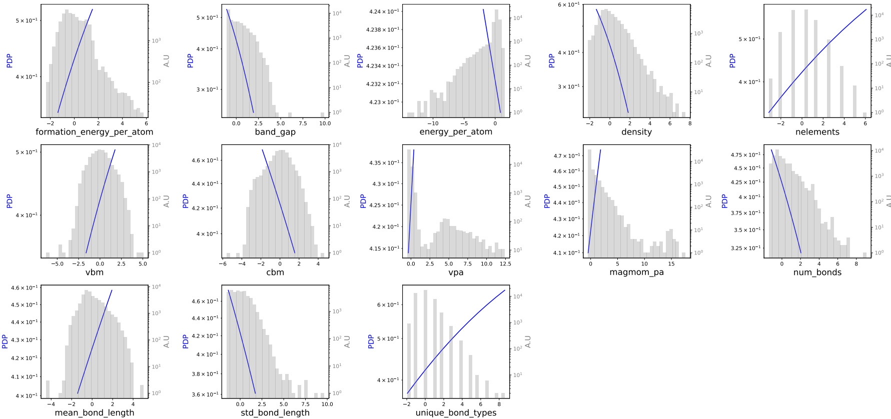
Partial dependence: A Visual technique used to understand the relationship between DNN prediction and a particular feature, while averaging out the effect of the rest of the features on the model probability.

Partial dependence plots help reveal how the model's prediction output varies across the range of a given feature, highlighting regions where the feature has a stronger or weaker influence.

Caveats:

- It works best with uncorrelated features
- As only the average probability is used, the spread in probability is ignored

Deep Neural Net: Partial Dependence Plots



PDP Feature Distributions

Deep Neural Net - Comparison

- expectations

Results

- The ML algorithms are trained to predict the thermodynamic stability by classifying the energy_above_hull feature. The energy_above_hull distribution is shown below.
- A cutoff of 0.05 eV/atom on energy_above_hull is used, as suggested by [4].

Summary & Future Goals

- Summary:
  - Overview to predict stability using ML for inorganic crystalline solid material
  - The role of thermodynamic stability in material design
- Future goals:
  - Moving beyond the prediction of stability to synthesizability

References

[1] Christopher J. Bartel. “Review of computational approaches to predict the thermodynamic stability of inorganic solids”. In: *Journal of Materials Science* 57.3 (2022), pp. 10478–10520. DOI: 10.1007/s10853-021-06865-6.

[2] Tianqi Chen and Carlos Guestrin. *XGBoost Documentation: Parameters*. https://xgboost.readthedocs.io/en/latest/parameter.html. Accessed: [Insert date]. 2023.

[3] B. Hao et al. “ElemNet: Deep Learning the Chemistry of Materials From Only Elemental Composition”. In: *arXiv preprint arXiv:1812.04153* (2018). URL: https://arxiv.org/abs/1812.04153.

[4] Geoffroy Hautier et al. “Data Mined Ionic Substitutions for the Discovery of New Compounds”. In: *Inorganic Chemistry* 50.2 (2011), pp. 656–663. DOI: 10.1021/ic100504j.

[5] Anubhav Jain et al. “Commentary: The Materials Project: A materials genome approach to accelerating materials innovation”. In: *APL Materials* 1.1 (2013), p. 011002. DOI:

Backup Slides

Magnetic moment per atom and volume per atom computation