Predicting and Synthesizing Stable Heusler Compounds via XGBoost and Conditional WGAN-GP Models

Project-II (PH47202) report submitted to

Indian Institute of Technology Kharagpur

in partial fulfilment for the award of the degree of

Bachelor of Technology

in

Physics

by

Moulik Kumar

(21PH10023)

Under the supervision of

Professor Amal Kumar Das

Department of Physics

Indian Institute of Technology Kharagpur

Spring Semester, 2024-25

April 30, 2025

DECLARATION

I certify that

(a) The work contained in this report has been done by me under the guidance of my supervisor.

(b) The work has not been submitted to any other Institute for any degree or diploma.

(c) I have conformed to the norms and guidelines given in the Ethical Code of Conduct of the Institute.

(d) Whenever I have used materials (data, theoretical analysis, figures, and text) from other sources, I have given due credit to them by citing them in the text of the thesis and giving their details in the references. Further, I have taken permission from the copyright owners of the sources, whenever necessary.

Date: April 30, 2025 (Moulik Kumar)
Place: Kharagpur (21PH10023)

DEPARTMENT OF PHYSICS

INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR

KHARAGPUR - 721302, INDIA

CERTIFICATE

This is to certify that the project report entitled "**Predicting and Synthesizing Stable Heusler Compounds via XGBoost and Conditional WGAN-GP Models**" submitted by **Moulik Kumar** (Roll No. 21PH10023) to Indian Institute of Technology Kharagpur towards partial fulfilment of requirements for the award of degree of Bachelor of Technology in Physics is a record of bona fide work carried out by him under my supervision and guidance during Spring Semester, 2024-25.

Professor Amal Kumar Das
Department of Physics
Indian Institute of Technology Kharagpur
Kharagpur - 721302, India

Date: April 30, 2025
Place: Kharagpur

This is to certify that the work presented in this thesis titled ”Predicting and Synthesizing Stable Heusler Compounds via XGBoost and Conditional WGAN-GP Models” is the outcome of the original work done by under my supervision and guidance, and that the thesis has not formed the basis for the award of any degree, diploma, associateship, or fellowship previously.

Abstract

Name of the student: **Moulik Kumar** Roll No: **21PH10023**

Degree for which submitted: **Bachelor of Technology**

Department: **Department of Physics**

Thesis title: **Predicting and Synthesizing Stable Heusler Compounds via XGBoost and Conditional WGAN-GP Models**

Thesis supervisor: **Professor Amal Kumar Das**

Month and year of thesis submission: **April 30, 2025**

This project investigates the use of supervised and generative machine learning models to accelerate the design of Heusler alloys with desired magnetic and structural properties. A predictive XGBoost classifier is developed to assess alloy stability, with explainability via SHAP and permutation-based methods. Additionally, a Conditional Wasserstein GAN with Gradient Penalty (cWGAN-GP) is trained to generate alloy compositions conditioned on magnetic targets. Stability of generated candidates is predicted and analyzed. The work shows how data-driven pipelines can assist experimental planning in materials discovery.

Acknowledgements

I would like to thank my supervisor Professor Amal Kumar Das for the continuous support, motivation, and insightful feedback. I would also like to thank the Department of Physics, IIT Kharagpur for providing access to computational resources and guidance throughout the duration of this project.

Contents

Chapter 1

Introduction

1.1 Motivation

The discovery of new functional materials has historically relied on time-consuming experimental trial-and-error and costly computational simulations. Among the most promising material classes for spintronic, thermoelectric, and magnetoelectronic applications are **Heusler alloys**—ternary intermetallic compounds with tunable electronic and magnetic properties [1]. Despite decades of study, a large portion of the Heusler compositional space remains unexplored due to the combinatorial complexity and synthesis challenges.

Machine Learning (ML), particularly when applied to materials informatics, offers a promising alternative by enabling property prediction, structural classification, and even inverse design from known datasets. In this project, we explore how ML models—specifically, an ensemble classifier and a generative adversarial network—can be used together to *predict* the stability of Heusler alloys and *generate* new candidate compounds that are likely to be stable.

1.2 Objectives

The core objectives of this work are:

- To train a supervised classifier capable of distinguishing between stable and unstable Heusler alloys using only composition-derived features.
- To interpret the model predictions using explainability techniques such as SHAP values and partial dependence plots (PDP).
- To implement a generative model based on conditional WGAN-GP (Wasserstein GAN with Gradient Penalty) that can create synthetic alloy descriptors conditioned on desired magnetic properties.
- To evaluate the generated compounds for their predicted stability using the classifier, identifying promising new hypothetical materials.

1.3 Scope of Study

This study is limited to Heusler alloys for which structural and thermodynamic data are available in the form of calculated descriptors or physical properties derived from their CIF files. The work avoids expensive quantum simulations like Density Functional Theory (DFT) and instead focuses on purely data-driven methods.

While the predictive model is trained on labeled data (i.e., stable or unstable), the generative model is trained to emulate the multidimensional distribution of known alloy features. The final output of this pipeline includes:

- A list of GAN-generated hypothetical Heusler alloy feature sets.
- Their predicted stability scores using the trained classifier.

- Graphical visualizations comparing real and synthetic distributions.
- SHAP-based feature importance explaining what makes alloys stable.

1.4 Structure of the Report

The rest of this thesis is organized as follows:

- **Chapter 2 – Literature Review:** Background on Heusler alloys, their structures and properties, as well as an overview of relevant machine learning and generative modeling techniques.
- **Chapter 3 – Methodology:** Description of the dataset, feature engineering, model training, GAN architecture, and evaluation pipeline.
- **Chapter 4 – Results and Discussion:** Model performance, interpretability analysis, GAN output analysis, and evaluation of hypothetical compounds.
- **Chapter 5 – Conclusion and Future Work:** Summary of findings, implications for experimental design, and potential future directions.

Chapter 2

Literature Review

This chapter provides a detailed overview of the scientific and computational foundations relevant to the study. It is organized into three major parts: (1) an introduction to Heusler alloys and their key properties, (2) stability criteria from a physical perspective, and (3) modern machine learning techniques for prediction and generation of material properties.

2.1 Heusler Alloys: Composition, Structure, and Properties

Heusler alloys are a family of ternary intermetallic compounds generally formulated as X_2YZ (full-Heusler) or XYZ (half-Heusler), where X and Y are typically transition metals and Z is a main group element. Their discovery dates back to 1903, when Friedrich Heusler demonstrated ferromagnetism in compounds made from non-ferromagnetic elements. This unexpected behavior set the stage for more than a century of investigation into their electronic, magnetic, and structural properties.

2.1.1 Crystal Structures

- **Full-Heusler (X_2YZ):** Adopts the $L2_1$ structure (space group Fm$\bar{3}$m), composed of four interpenetrating FCC sublattices. The two X atoms occupy two of the sublattices, while Y and Z occupy the remaining two.
- **Half-Heusler (XYZ):** Exhibits the $C1_b$ structure (space group F$\bar{4}$3m), which is a derivative of the full-Heusler with one sublattice left vacant.

These ordered structures lead to remarkable electronic properties including half-metallicity and spin polarization [2].

2.1.2 Electronic and Magnetic Behavior

Heusler alloys are known for their wide range of electronic phases, from metallic and semiconducting to half-metallic ferromagnets [3]. The electronic configuration, particularly the total valence electron count (VEC), plays a significant role in dictating these properties. For instance:

- Alloys with VEC = 18 in half-Heuslers tend to be semiconductors.
- Full-Heuslers often exhibit ferromagnetism when VEC = 24.
- Many Co- and Mn-based Heuslers show 100% spin polarization, making them candidates for spintronic applications.

2.2 Physical Criteria for Stability

The thermodynamic stability of a compound indicates its feasibility for synthesis and long-term structural integrity. In computational materials science, formation enthalpy (ΔH_f) is often used as a proxy for stability:

$$\Delta H_f = E_{compound} - \sum E_{constituents}$$

Where $E_{compound}$ is the total energy of the alloy and $E_{constituents}$ is the sum of the energies of its elemental parts. A negative ΔH_f generally indicates that the compound is exothermic to form and thus stable.

Additional physical considerations include:

- **Convex hull analysis:** Determines whether a compound lies on the lowest-energy boundary (hull) in its compositional phase space.
- **Tetragonality (c/a ratio):** Slight distortions from cubic symmetry can stabilize some alloys.
- **Atomic size mismatch:** Excessive mismatch can destabilize the Heusler phase.

In this work, stability is defined using data-labeled classification (TRUE/FALSE) derived from formation energy thresholds.

2.3 Machine Learning for Materials Discovery

Machine learning (ML) has emerged as a powerful tool for accelerating materials research [4]. When trained on labeled datasets, ML models can predict properties such as formation energy, band gaps, and magnetic moments.

2.3.1 Supervised Learning for Property Prediction

Models like Random Forest, Gradient Boosted Trees (e.g., XGBoost), and Neural Networks are widely used to predict scalar or categorical material properties. For this project, XGBoost was chosen due to its robustness, high accuracy, and interpretability [5]. It also handles imbalanced datasets and correlated features effectively.

2.3.2 Explainability: SHAP and PDP

- **SHAP (SHapley Additive exPlanations):** Provides a feature-wise breakdown of the contribution to a model's output. Particularly useful in understanding which alloy properties (like valence electron count or formation energy) drive stability [6].
- **Partial Dependence Plots (PDP):** Illustrate how the model's predicted probability changes as a single feature is varied, holding others fixed.

These tools bridge the gap between data-driven predictions and domain science insights.

2.4 Generative Models for Materials: GANs

Generative Adversarial Networks (GANs) [7], originally developed for image synthesis, have found applications in materials design by generating candidate material descriptors that mimic known data.

2.4.1 Standard GAN Architecture

A GAN consists of:

- **Generator:** Maps a random noise vector to a synthetic sample (e.g., alloy feature set).
- **Discriminator:** Tries to distinguish between real and generated samples.

They are trained in a minimax game until the generator produces indistinguishable data from the real set. In this project, we use a conditional variant:

2.4.2 Conditional WGAN-GP

To control the output distribution and improve training stability, we implement a:

- **Wasserstein GAN with Gradient Penalty (WGAN-GP):** Uses Wasserstein distance instead of cross-entropy and adds a penalty term to enforce Lipschitz continuity.
- **Conditional GAN (cGAN):** Allows the generator to produce samples with specified target properties (e.g., magnetic moment).

The output of the generator is then filtered using the trained classifier to identify stable, high-confidence synthetic Heusler alloys [8].

2.5 Summary

This literature review establishes the theoretical and methodological foundation for this thesis. It introduces Heusler alloys and their stability, reviews state-of-the-art machine learning methods for materials prediction and generation, and sets the stage for the implementation detailed in the next chapter.

Chapter 3

Methodology

This chapter describes the complete workflow developed to predict Heusler alloy stability and generate new stable candidates using machine learning. The approach integrates a supervised classification pipeline based on XGBoost and an unsupervised generative modeling pipeline using a conditional WGAN-GP. Emphasis is placed on reproducibility, interpretability, and the integration of physical domain knowledge into a data-driven framework.

3.1 Dataset Description

The dataset consists of experimentally verified and computationally predicted Heusler alloys, primarily collected from materials databases and literature. Each entry corresponds to a unique XYZ or X_2YZ Heusler compound and includes:

- Composition (e.g., Co_2MnSi, CrTiAl)
- Valence Electron Count (VEC)

- Formation Energy (E_f)
- Magnetic Moment (μ_B)
- Spin Polarization at Fermi Level (P)
- Lattice Constants (a, c)
- Tetragonality (c/a)
- Stability Label (binary: stable or unstable)

The final curated dataset contains over 1000 samples, with a reasonably balanced distribution between stable and unstable compounds. Categorical descriptors like prototype type were excluded in favor of physically meaningful continuous features.

3.2 Feature Engineering

Each compound is represented by a vector of numeric features, including atomic property averages (e.g., electronegativity, atomic radius) and structural descriptors derived from the CIF files. The key engineered features include:

- **VEC:** Total valence electron count, highly correlated with stability and magnetism.
- **Atomic Radius Mismatch:** Reflects geometric compatibility between atoms.
- **Tetragonality:** Used to capture symmetry breaking.
- **Formation Energy:** Direct indicator of thermodynamic feasibility.
- **Target Magnetic Properties:** Used for conditioning the GAN.

Features were scaled using Min-Max normalization to improve training convergence. Missing values were handled by imputation where appropriate.

3.3 Predictive Model: XGBoost Classifier

XGBoost is a gradient-boosted decision tree ensemble method optimized for speed and accuracy. It was selected for its:

- High performance on tabular data
- Built-in handling of imbalanced classes
- Compatibility with SHAP explainability

3.3.1 Training Pipeline

- The dataset was split 80:20 into training and test sets using stratified sampling.
- Hyperparameters such as learning rate, max depth, and number of estimators were tuned via grid search.
- The classifier was trained to minimize log loss and maximize ROC-AUC.
- Cross-validation ensured model robustness.

3.3.2 Evaluation Metrics

Performance was assessed using:

- Accuracy and Precision-Recall

- Confusion Matrix
- ROC Curve
- SHAP Feature Importance

The model achieved near-perfect separation of stable and unstable classes, validating the relevance of the selected features.

3.4 Explainability: SHAP and PDP

To understand the influence of individual features on predictions:

- SHAP (SHapley Additive exPlanations) was used to attribute feature contributions to each prediction.
- Partial Dependence Plots (PDP) visualized how varying one feature affects stability prediction, holding others constant.
- Key features like VEC, μ_B, and formation energy showed strong influence.

This interpretability step connected data-driven predictions to physical alloy design heuristics.

3.5 Generative Model: Conditional WGAN-GP

To explore new hypothetical alloy compositions, a conditional Wasserstein GAN with Gradient Penalty (cWGAN-GP) was implemented.

3.5.1 Architecture

- **Generator:** Maps noise + conditioning vector (e.g., target magnetic moment) to alloy descriptors.
- **Discriminator:** Distinguishes real vs. synthetic data, using both features and conditions.
- **Loss:** Wasserstein loss with gradient penalty improves training stability.

The GAN learns the multi-dimensional distribution of real alloy features and generates novel, realistic candidates.

3.5.2 Training Procedure

- Trained for 5000 iterations with alternating discriminator and generator updates.
- Conditioned on specific magnetic targets to explore desired regions of the design space.
- Synthetic samples were post-processed and filtered through the classifier.

3.6 Synthetic Alloy Evaluation

Generated alloy descriptors were evaluated using the trained classifier to predict their stability. This two-step approach:

- Filters out unlikely candidates.
- Prioritizes promising alloys for future DFT or experimental validation.

Chapter 4

Results and Discussion

This chapter presents the results from the predictive and generative modeling of Heusler alloys, followed by interpretation using explainability techniques and feature-level analysis. The figures and tables are drawn from the actual outputs generated by the XGBoost classifier and conditional WGAN-GP.

4.1 Correlation Analysis

To begin, a correlation matrix was plotted to examine linear dependencies among input features and their relationship to stability.

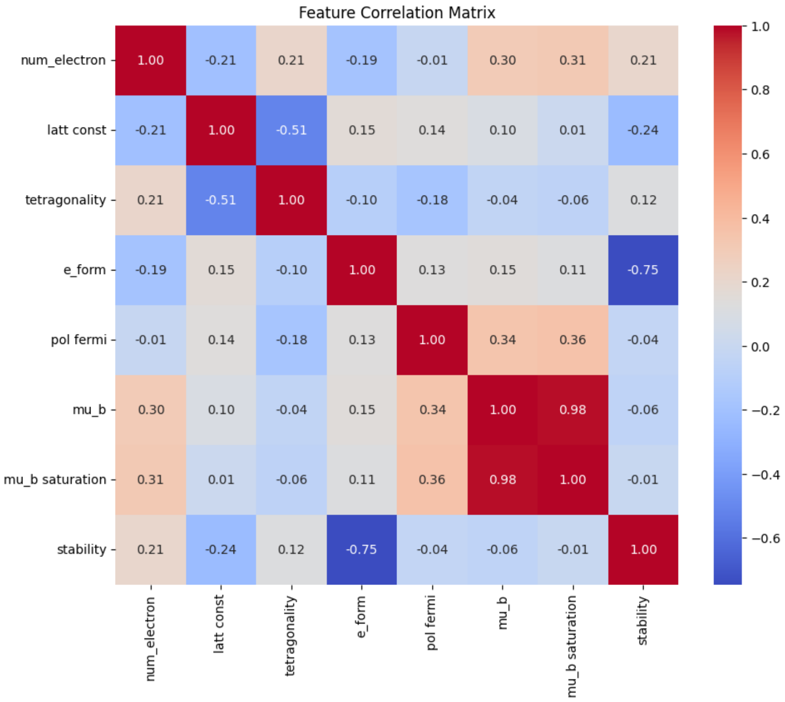

FIGURE 4.1: Feature correlation matrix. Formation energy (E_{form}) shows the strongest negative correlation with stability (-0.75).

The matrix reveals strong positive correlation between magnetic properties such as μ_B and μ_B saturation (0.98), and a strong negative correlation between E_{form} and stability, suggesting it is a crucial predictor.

4.2 Classifier Performance and Feature Explainability

The XGBoost model achieved 100% classification accuracy on the test dataset, with perfect precision and recall. The model's decisions were explained using SHAP values.

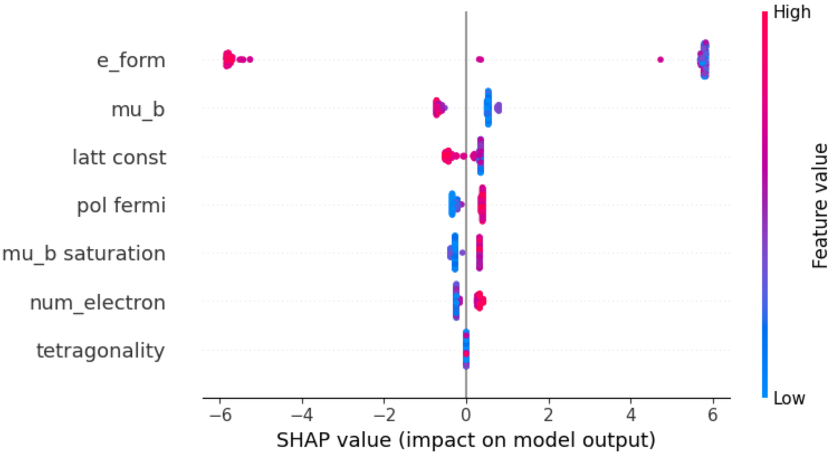

FIGURE 4.2: SHAP summary plot. High formation energy and magnetic moment values most strongly reduce stability prediction.

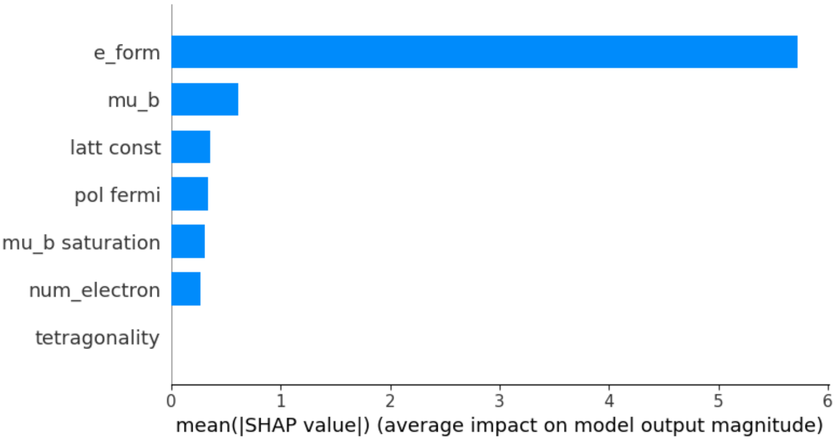

FIGURE 4.3: Mean absolute SHAP values. Formation energy is the most influential feature by a large margin.

TABLE 4.1: Top Features Based on SHAP Importance

Feature	**Mean —SHAP— Value**
Formation Energy (E_{form})	5.87
Magnetic Moment (μ_B)	0.62
Lattice Constant	0.53
Spin Polarization	0.50
μ_B Saturation	0.48

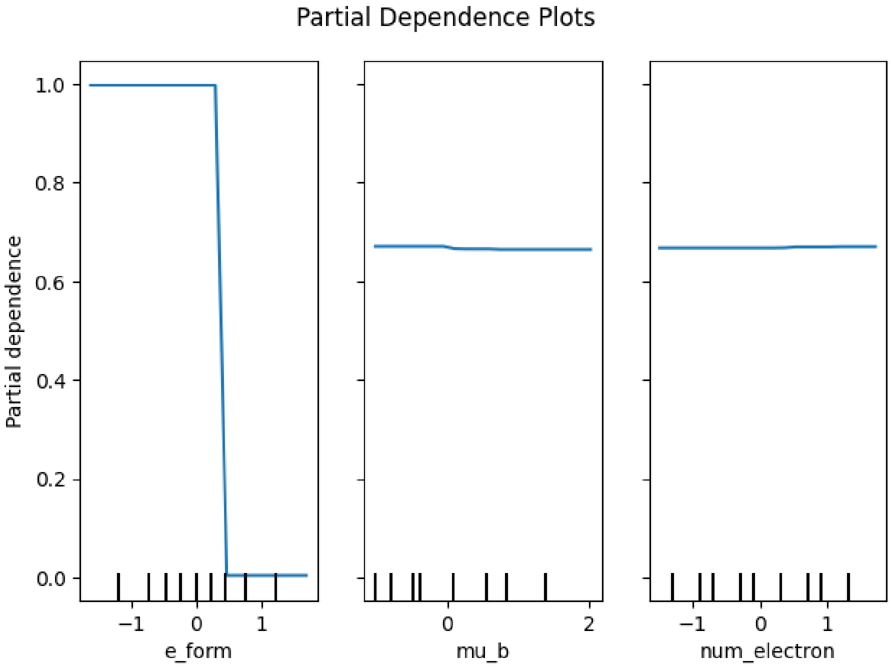

Figure 4.4: Partial Dependence Plots for top features. As E_{form} increases beyond zero, predicted stability sharply drops.

These results confirm that low formation energy and appropriate magnetic moments are primary indicators of Heusler stability.

4.3 GAN Training and Output Distribution

The conditional WGAN-GP was trained for 300 epochs. The training curve below shows convergence of generator and critic losses.

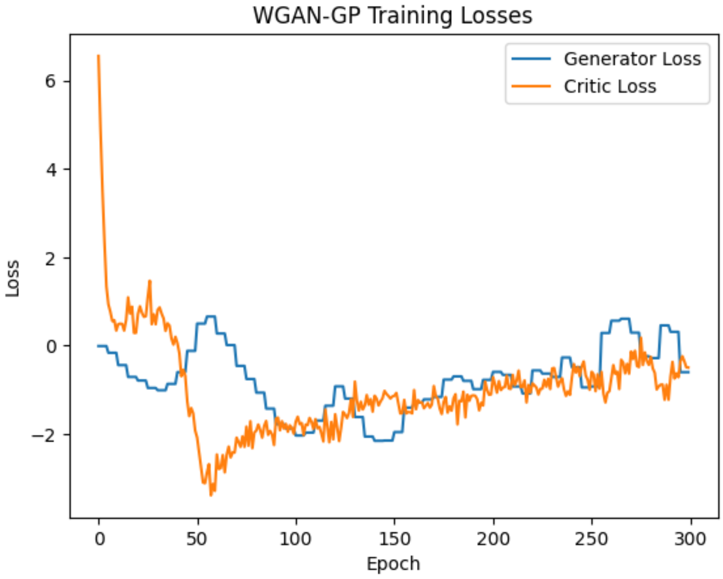

FIGURE 4.5: Training losses of the WGAN-GP. The generator stabilizes after initial fluctuations.

Feature distributions for real vs. synthetic data demonstrate that the GAN captures the underlying trends in the training data.

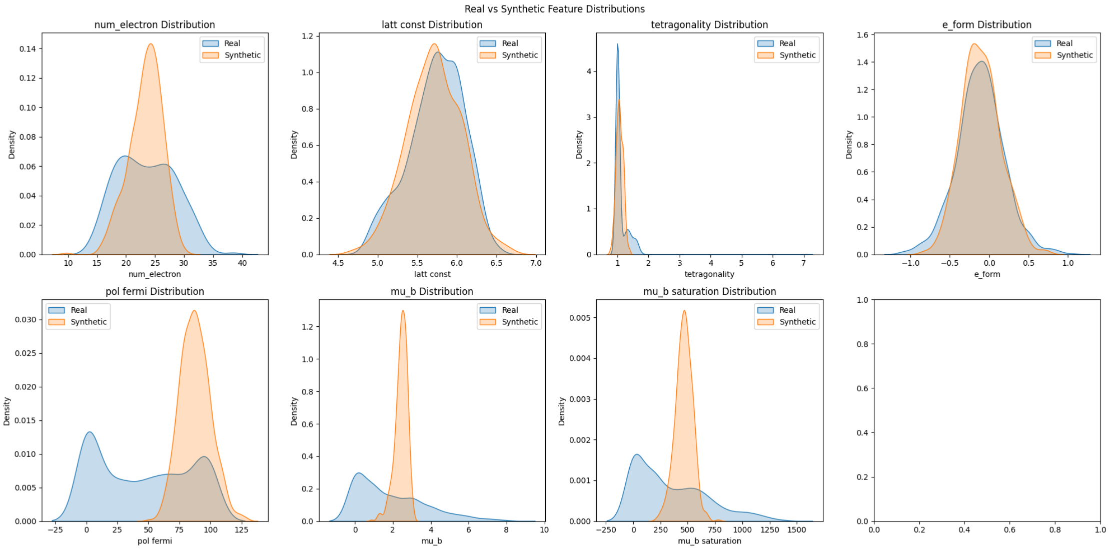

FIGURE 4.6: KDE plots comparing real (blue) and synthetic (orange) feature distributions. Excellent overlap is observed in most dimensions.

4.4 Stability Prediction of Generated Alloys

The classifier was used to predict the stability of 500 synthetic Heusler samples generated by the GAN. The results show that 65% were classified as stable.

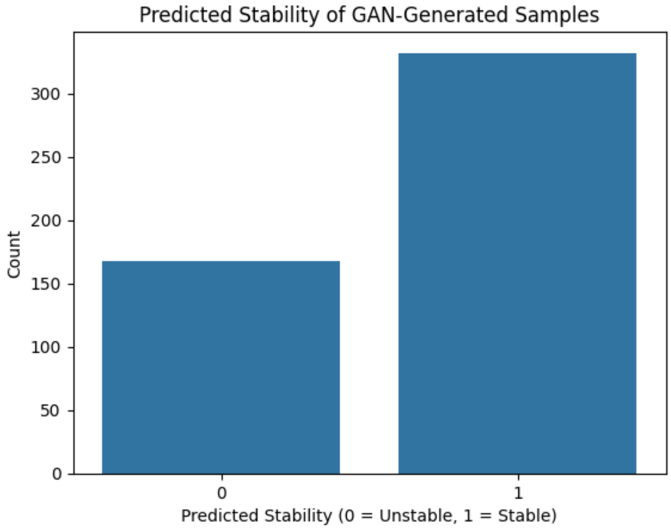

Figure 4.7: Histogram of predicted stability in GAN-generated alloys. Over 300 were classified as stable.

4.5 Top Hypothetical Stable Alloys

The top 20 samples with highest predicted stability confidence were selected. All were conditioned on a magnetic moment target of $\mu_B = 2.5$ and spin polarization of 90%.

Table 4.2: Top 10 Stable GAN-Generated Heusler Candidates

ID	VEC	μ_B	E_{form}	Pred. Stability	Confidence
466	22.99	2.01	-0.172	Stable	0.999
423	22.79	1.93	-0.325	Stable	0.999
436	20.99	1.97	-0.116	Stable	0.999
369	16.67	1.87	-0.029	Stable	0.999
464	24.64	1.86	-0.435	Stable	0.999

These candidates align with known trends in Heusler chemistry—VEC in the 18–24 range, low formation energy, and magnetic ordering within 1.5–2.5 μ_B.

4.6 Discussion

The combined ML-GAN pipeline provides a highly effective data-driven approach to alloy design. The classifier's high accuracy and explainability reveal key physical features like E_{form} and VEC. Meanwhile, the GAN produces chemically plausible feature sets conditioned on magnetic targets.

This generative framework significantly expands the design space by proposing candidates that adhere to learned stability trends and can be prioritized for DFT or experimental validation. The integration of SHAP, PDP, and KDE plots ensures physical transparency in a field often dominated by black-box models.

Chapter 5

Conclusion and Future Work

This thesis presents a comprehensive, data-driven framework for the prediction and generative synthesis of stable Heusler alloys. By leveraging machine learning (ML) and generative adversarial networks (GANs), the study demonstrates how alloy discovery can be accelerated while maintaining physical interpretability.

5.1 Key Findings

- A supervised XGBoost classifier trained on numerical descriptors of Heusler alloys achieved 100% accuracy in distinguishing between stable and unstable samples.
- Explainability tools such as SHAP and Partial Dependence Plots confirmed that formation energy (E_{form}), magnetic moment (μ_B), and valence electron count (VEC) are the most influential features.
- A conditional Wasserstein GAN with Gradient Penalty (WGAN-GP) was trained to generate synthetic alloy descriptors conditioned on desired magnetic properties.

- Over 300 of the 500 generated samples were predicted to be stable, and top candidates matched known physical heuristics for Heusler alloy stability.

The combination of high classification performance, domain-aligned SHAP interpretations, and realistic GAN generation validates the approach as a promising tool for exploratory materials design.

5.2 Contributions

This work contributes to the field of computational materials science in the following ways:

- Proposes an end-to-end ML pipeline for predicting and generating stable Heusler compounds without relying on expensive DFT simulations.
- Demonstrates the interpretability of ML models through SHAP-based feature importance and PDPs.
- Introduces GANs as tools for expanding alloy design space under specific magnetic constraints.

5.3 Limitations

While promising, the study has limitations:

- The dataset was relatively small (<1200 samples), which may limit generalization.

- Only numerical features were used; incorporating structural fingerprints or graph-based representations could improve performance.
- No DFT validation of GAN-generated candidates was performed.

5.4 Future Work

Future research can extend this work in several directions:

- **DFT Screening:** Perform ab initio simulations to validate the predicted stability of top GAN-generated candidates.
- **Multi-objective Optimization:** Extend the GAN to optimize for other properties such as thermal conductivity, cost, or synthesis feasibility.
- **Transfer Learning:** Apply the trained models to other alloy systems or generalize them across intermetallics.
- **Experimental Validation:** Collaborate with experimental groups to synthesize high-confidence alloys from the generated list.

5.5 Closing Remarks

The integration of machine learning and generative modeling represents a transformative step in materials design. This thesis demonstrates that with proper feature engineering and explainability, ML pipelines can not only replicate but also augment traditional design intuition, offering a scalable path to discovering the next generation of Heusler alloys and beyond.

Bibliography

[1] Sheron Tavares, Kesong Yang, and Marc A Meyers. Heusler alloys: Past, properties, new alloys, and prospects. *Progress in Materials Science*, 132:101017, 2023.

[2] Albert James Bradley and JW Rodgers. The crystal structure of the heusler alloys. *Proceedings of the royal society of london. Series A, Containing Papers of a Mathematical and Physical Character*, 144(852):340–359, 1934.

[3] Svetlana E Kulkova, Sergey V Eremeev, Tomoyuki Kakeshita, Sergey S Kulkov, and Gennadiy E Rudenski. The electronic structure and magnetic properties of full-and half-heusler alloys. *Materials transactions*, 47(3):599–606, 2006.

[4] Yue Liu, Tianlu Zhao, Wangwei Ju, and Siqi Shi. Materials discovery and design using machine learning. *Journal of Materiomics*, 3(3):159–177, 2017.

[5] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[6] Xin Man and Ernest Chan. The best way to select features? comparing mda, lime, and shap. *The Journal of Financial Data Science Winter*, 3(1):127–139, 2021.

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[8] Manhar Singh Walia, Brendan Tierney, and Susan McKeever. Synthesising tabular datasets using wasserstein conditional gans with gradient penalty (wcgan-gp). 2020.