ACCELERATING ALLOY DESIGN: PREDICTING GRAIN BOUNDARY SEGREGATION VIA DATA-DRIVEN MODELS

Presented by (GROUP 10)-

BAKARE SHIVRAJ SHIVAJI 230280
BASUDEV MOHAPATRA 230286
KHUSHAL KARADIYA 230558
PRUTHA VINAY 230807
GAURAV KUMAR 230411

Problem Statement

The project's goal is to predict grain boundary (GB) segregation energies in polycrystalline alloys using a machine learning and deep learning pipeline.

Grain boundaries, the interfaces between crystal grains in a material, critically influence its most important properties, including mechanical strength, corrosion resistance, and thermal stability.

The core problem is that traditional methods for calculating these segregation energies, such as Density Functional Theory (DFT) simulations, are computationally prohibitive. They are too slow and resource-intensive to be used for large-scale screening of new potential alloys.

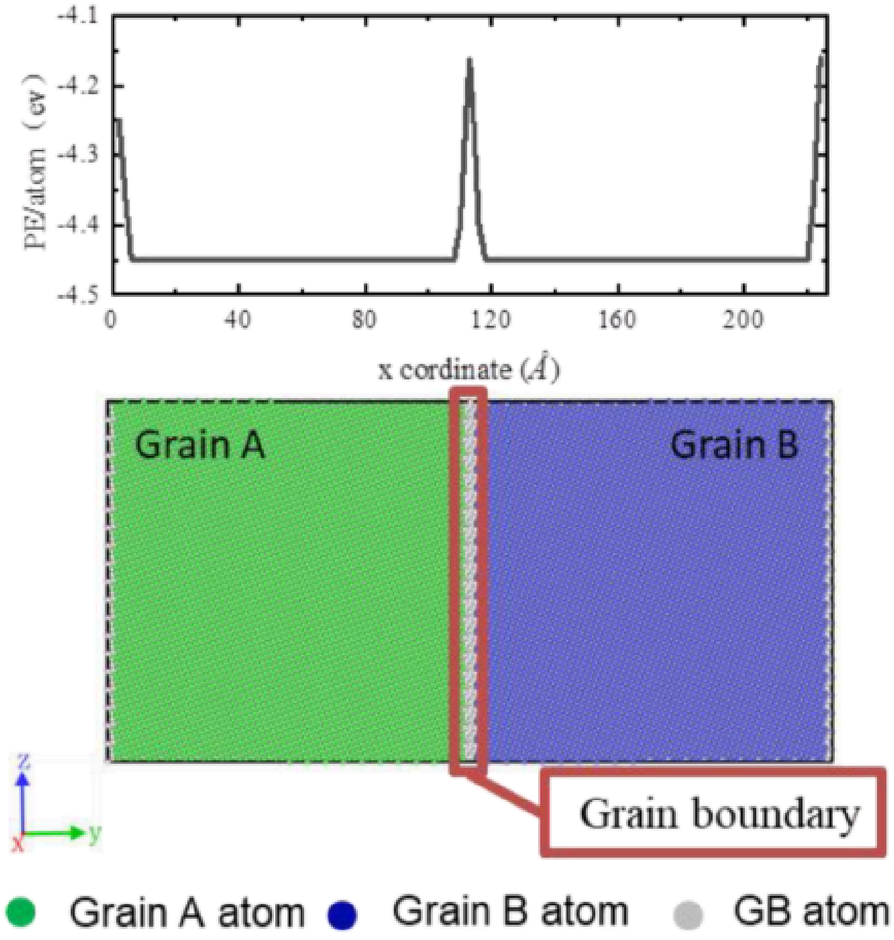
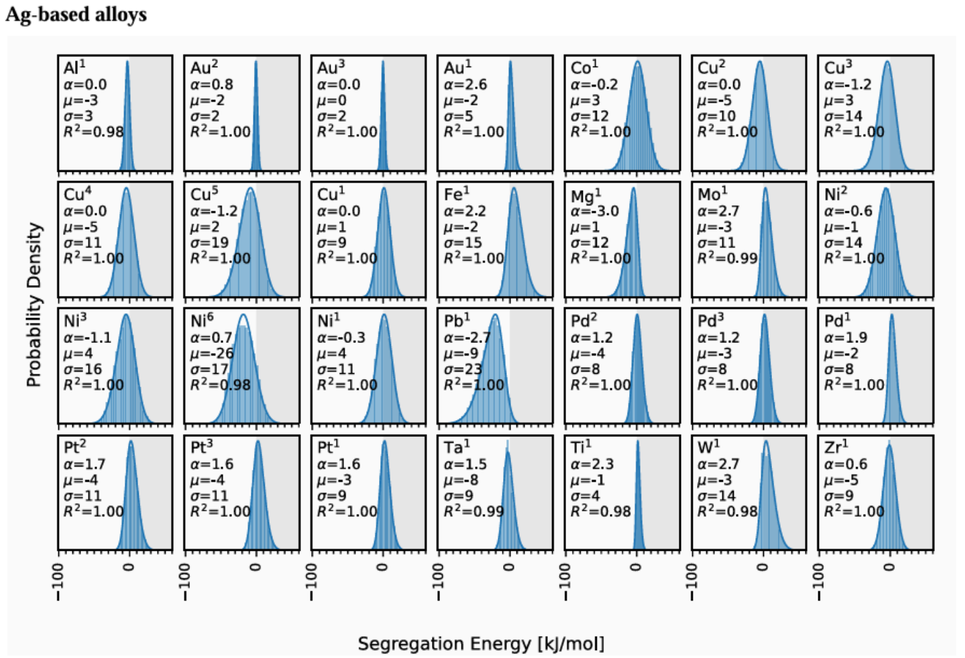


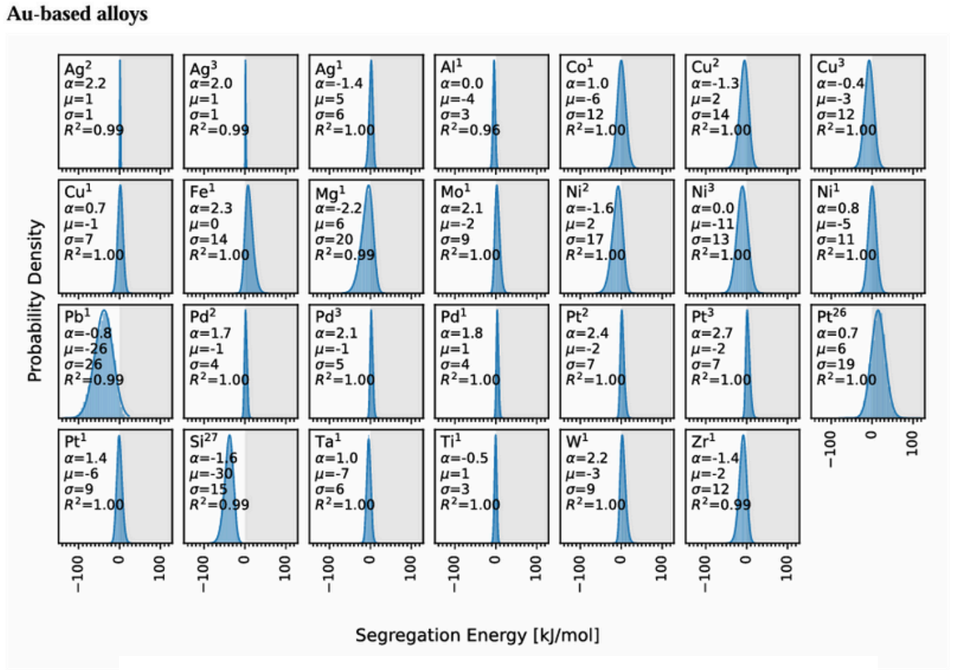
Fig 1: PE/atom V/s x coordinate

Therefore, this project aims to build a data-driven model that learns the complex relationship between a material's fundamental properties and its segregation behavior.

The core theory is that segregation is driven by atomic and elastic mismatch (differences in density, stiffness, etc.) between solvent and solute atoms. This energy isn't a single value but a statistical distribution defined by parameters like mean, width, and skewness.

We use traditional ML models (like Linear Regression and Random Forest) on engineered mismatch features and Deep Learning (like GNNs) on atomic structures to learn this complex physics-based relationship. The goal is to create a fast, accurate model for screening new alloys.





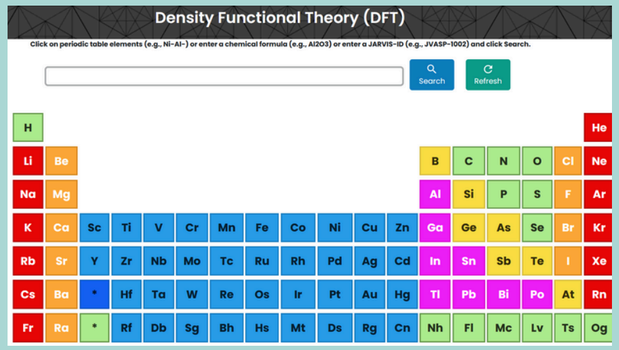
METHODOLOGIES: DATA COLLECTION

Data Required

TARGET DATA: Mean Segregation Energy (μ)

FEATURE DATA: Fundamental bulk properties of the pure metals that drive segregation

JARVIS-DFT



Materials Project (MP) API

All **154,879 materials**
Showing 1-15

Columns | Export Table

Material ID	Formula	Crystal System	Space Group Symbol	Sites	Energy Above Hull (eV/atom)	Band Gap (eV)
mp-862690	Ac	Hexagonal	$P6_3/mmc$	4	0	0
mp-1183377	Ac_2InGa	Cubic	$Fm\bar{3}m$	4	0	0
mp-1183093	Ac_2IrRh	Cubic	$Fm\bar{3}m$	4	0	0
mp-867241	Ac_2GePd	Cubic	$Fm\bar{3}m$	4	0	0

Retrieval

The Problem Data - Atomistic data quantifying segregation in binary alloys

	A	B	C	D	E	F	G
1	Solvent	Solute	Potential_	Alpha	Mu	Sigma	R_Squared
2	Ag	Al	1	0	-3	3	0.98
3	Ag	Au	2	0.8	-2	2	1
4	Ag	Au	3	0	0	2	1
5	Ag	Au	1	2.6	-2	5	1
6	Ag	Co	1	-0.2	3	12	1
7	Ag	Cu	2	0	-5	10	1
8	Ag	Cu	3	-1.2	3	14	1

Initial Feature Set - (ρ,K,G) for 6 FCC metals (Al, Cu, Ag, Au, Pt, Pd) from JARVIS, MP

Gap Filler- Used a live API query for the missing feature data **(>80%)** of dataset

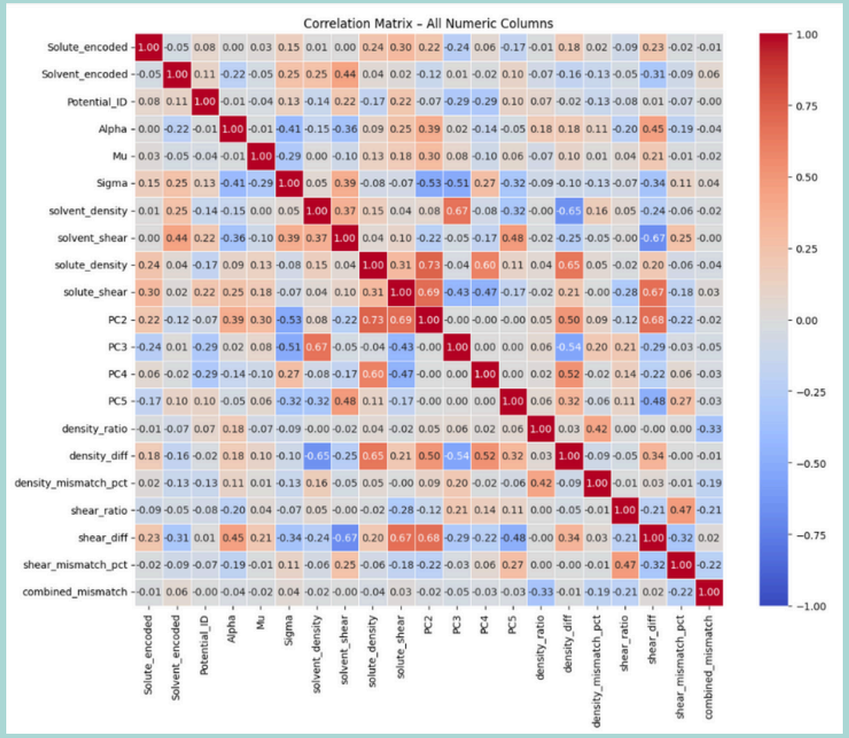
```
from mp_api.client import MPRester
import pandas as pd

materials = ["Al", "Cu", "Au", "Ag", "Pt", "Pd"]
records = []

with MPRester("eX5BKhk5dTLiG861eZ8Y2iO7gJpp3mdY") as mpr:
    for el in materials:
        entries = mpr.materials.summary.search(
            elements=[el],
            fields=["material_id", "formula_pretty", "formation_energy_per_
        )
        for e in entries:
            records.append({
                "material_id": e.material_id,
                "formula": e.formula_pretty,
                "formation_energy_per_atom": e.formation_energy_per_atom
            })
```

Pre-processing and Feature Engg.

INITIAL FILTERING - Keep only the most stable ground state
NORMALIZING - Apply std. scaling to numerical features
CATEGORICAL ENCODING - Convert element names to numerical values for ML model to process
MISMATCH FEATURES - eg. Shear Ratio, Density Difference
COMPOSITE FEATURES - Couple multiple properties eg. Density and Shear modulus

Solute_en	Solvent_e
1	0
2	0
2	0
2	0
3	0
5	0

shear_mismatc	combined_mism
-0.277426573	-6.24664
-0.197870764	4.765797
-0.197870764	4.765797
-0.197870764	4.765797
-1.457903139	5.590073
-0.830395519	3.303397
-0.830395519	3.303397
-0.830395519	3.303397
-0.830395519	3.303397



Predictive Regression Modeling

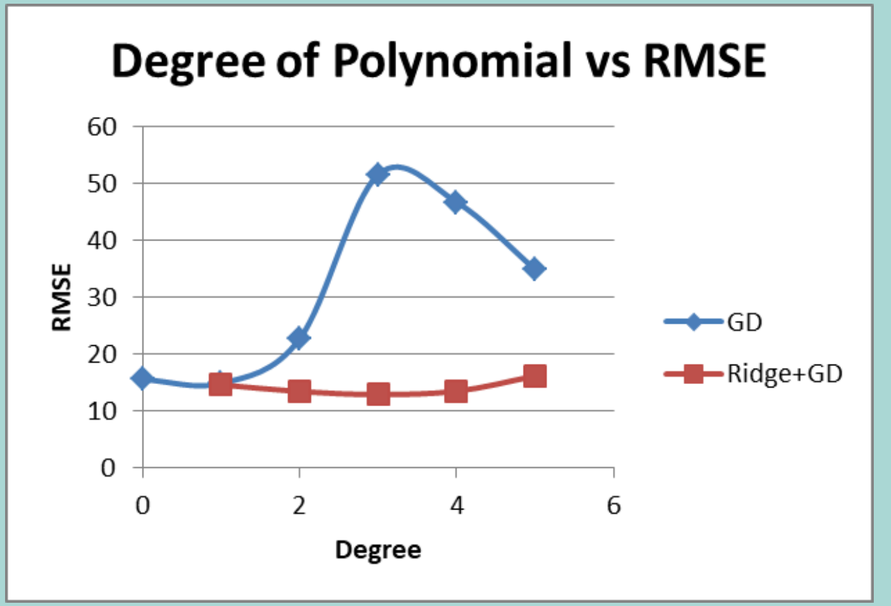
Technique	Purpose
• Polynomial Regression Model	Models complex non-linear patterns that linear regression cannot capture.
• Gradient Descent Optimization	Iteratively updates weights to minimize error via gradient descent.
• Ridge Regression (L2)	Regularization to avoid overfitting and ensure stable convergence.
• Grid Search	Finds the optimal model degree and L2 regularization parameter (α).
• Cross Validation	Ensures reliable model validation and avoids biased evaluation.

LIBRARIES USED : 📚 NUMPY 📊 PANDAS 📈 MATPLOTLIB ⚙️ SCIKIT-LEARN 💾 OS

Prediction and Error Analysis

Performance Summary

Model	Degree	Best(alpha)	Test Rmse
Gradient	1	-	15.6
Ridge+GD	3	3.16	12.9

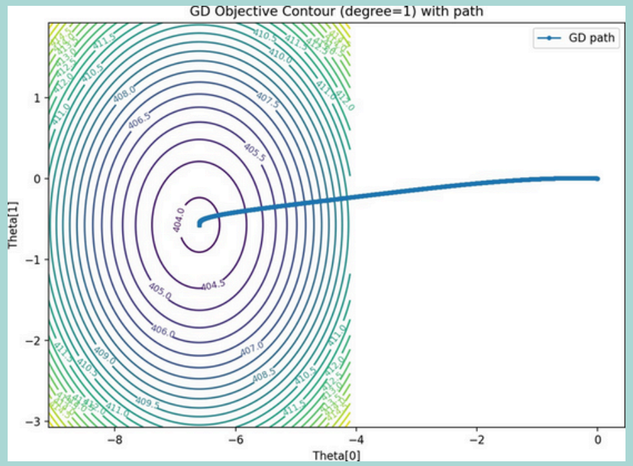


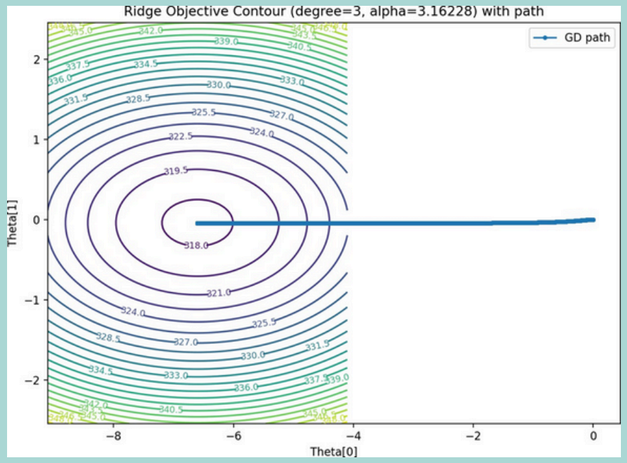
Insights

- Low-degree → underfitting; high-degree → overfitting
- L2 regularization stabilizes training & reduces test error
- Gradient descent path visualized on contour confirms convergence

Visuals

GD Contour Plot





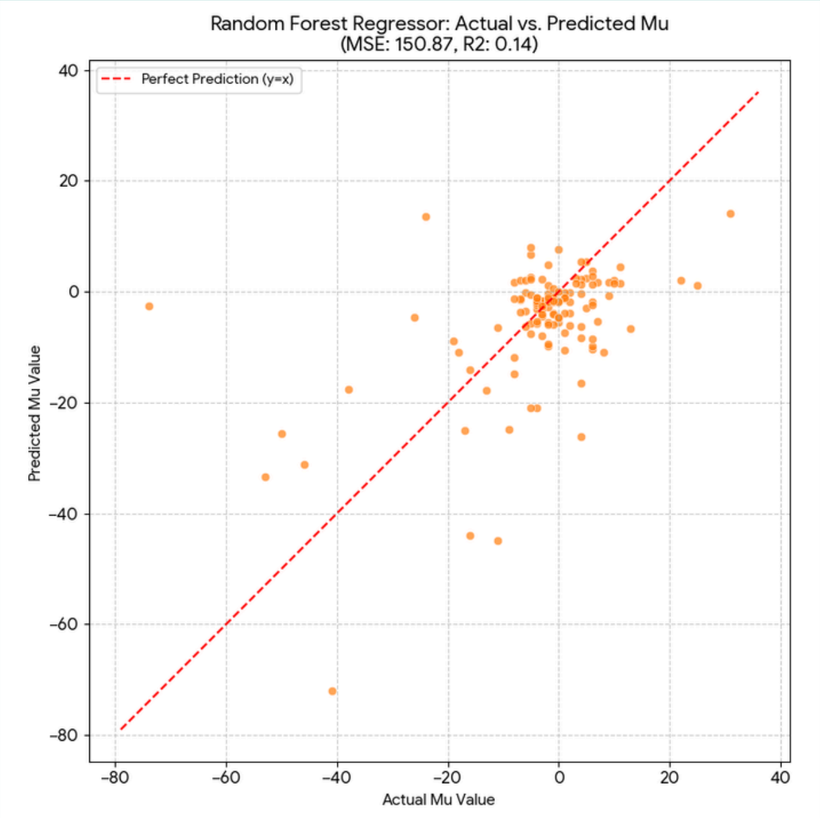
Ridge Contour Plot

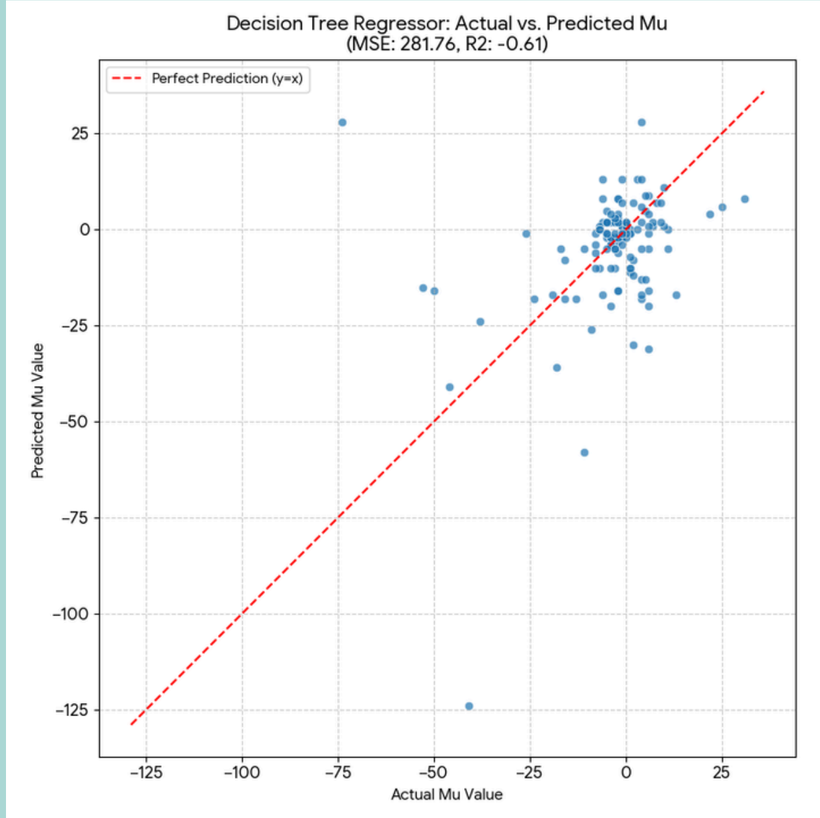
Modelling of Decision Tree Based on Regression

Model	Description	Primary Goal
Decision Tree Regressor (DT)	A single, non-linear model.	Establishing a baseline performance metric.
Random Forest Regressor (RF)	An ensemble of 100 Decision Trees (Bagging).	Mitigating the high variance and overfitting characteristic of a single Decision Tree.

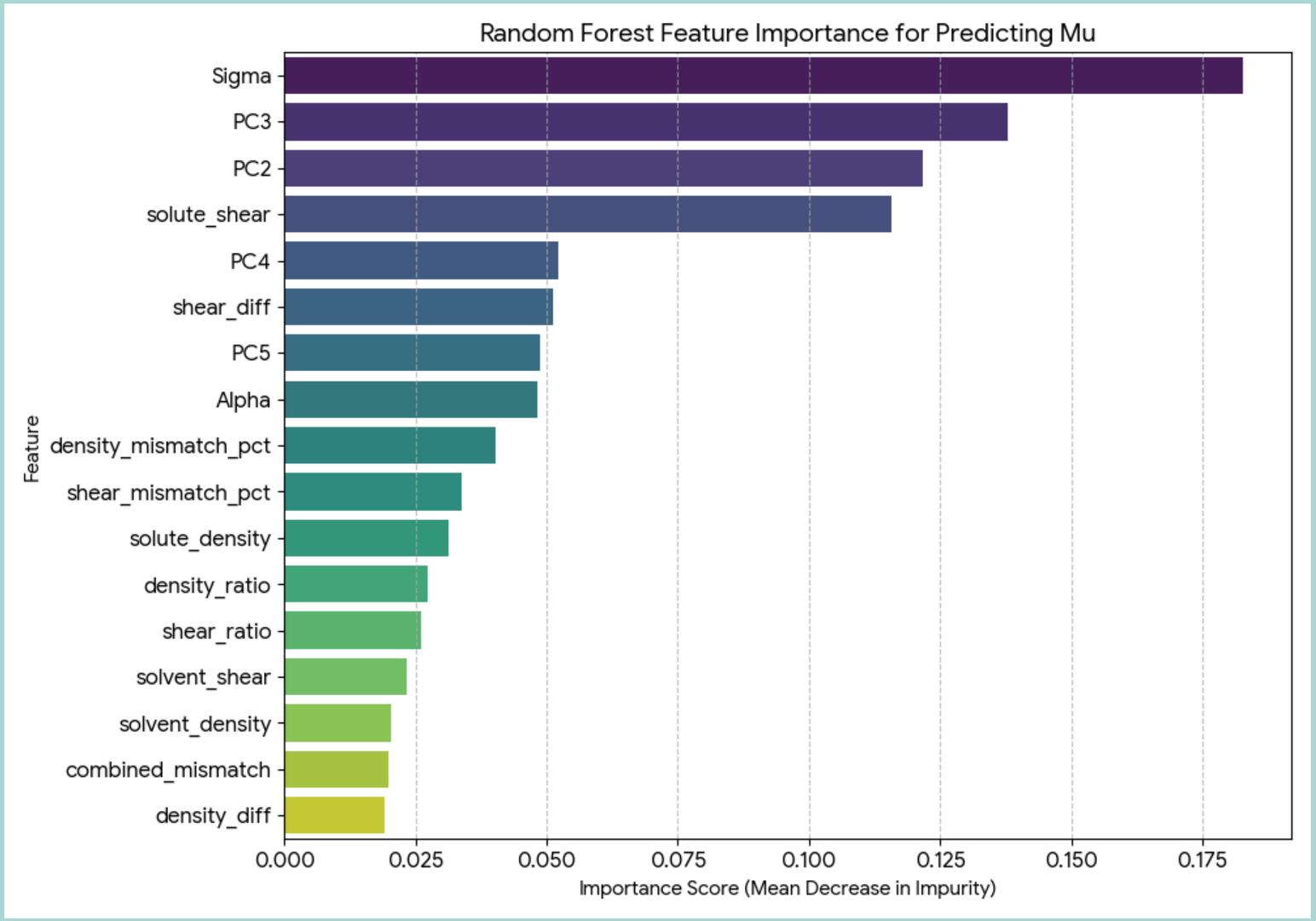
Model	Mean Squared Error (MSE)	R-squared (R2) Score
Decision Tree Regressor	281.76	-0.6127
Random Forest Regressor	150.87	0.1365

Prediction and Error Analysis





- **Decision Tree:** The points are highly scattered and far from the perfect prediction line, especially at the extremes, confirming the poor generalization and negative R^2.
- **Random Forest:** The scatter is visibly tighter than the Decision Tree, but a large degree of error remains. The predicted values are centered around the perfect line but lack precision, leading to the low positive R^2 score. This scatter highlights the inherent difficulty in predicting 'Mu' with the current feature set.

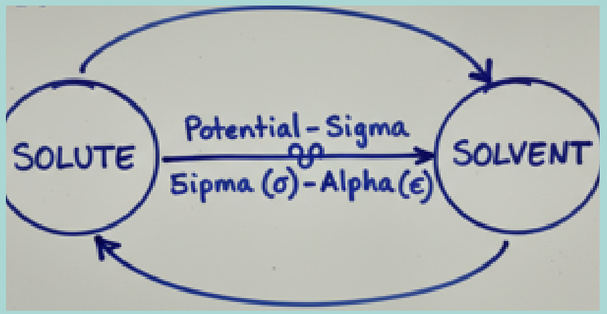


This graph shows the model's predictions are scattered around the perfect prediction line (y=x), resulting in a low R^2 score of 0.14. While the ensemble model reduced the error significantly compared to the Decision Tree, the substantial scatter indicates that it lacks the precision to reliably predict 'Mu' and suggests the features are not sufficiently predictive.

Predicting Physical Properties using Graph Neural Networks

Introduction

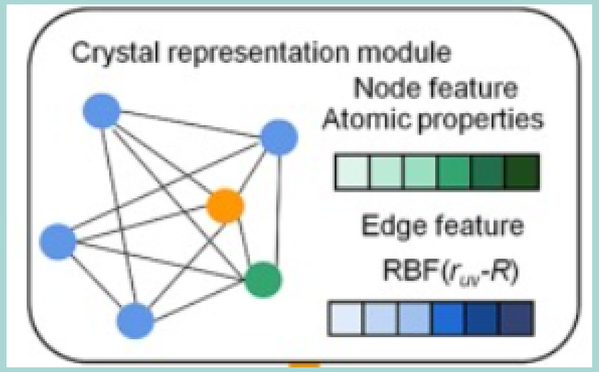
- Traditional regression models fail to capture relational/graph-based interactions between solute and solvent
- Physical properties like viscosity (μ) depend on interactions rather than individual attributes



- GNNs can model these relationships efficiently

GIN

- The GIN is designed to achieve powerful graph-level representations by aggregating neighboring node features in a way that maximizes discriminative power

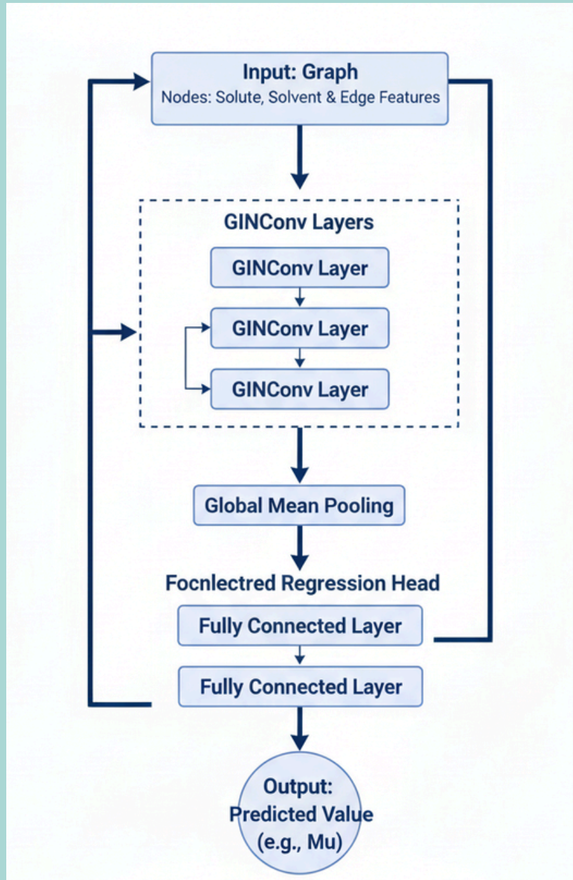


- The update rule for each node v at the k-th layer is given by:

$$h_v^{(k)} = MLP^{(k)}\left((1+\varepsilon^{(k)})h_v^{(k-1)} + \sum_{u\in\mathcal{N}(v)} h_u^{(k-1)}\right)$$

Model Architecture

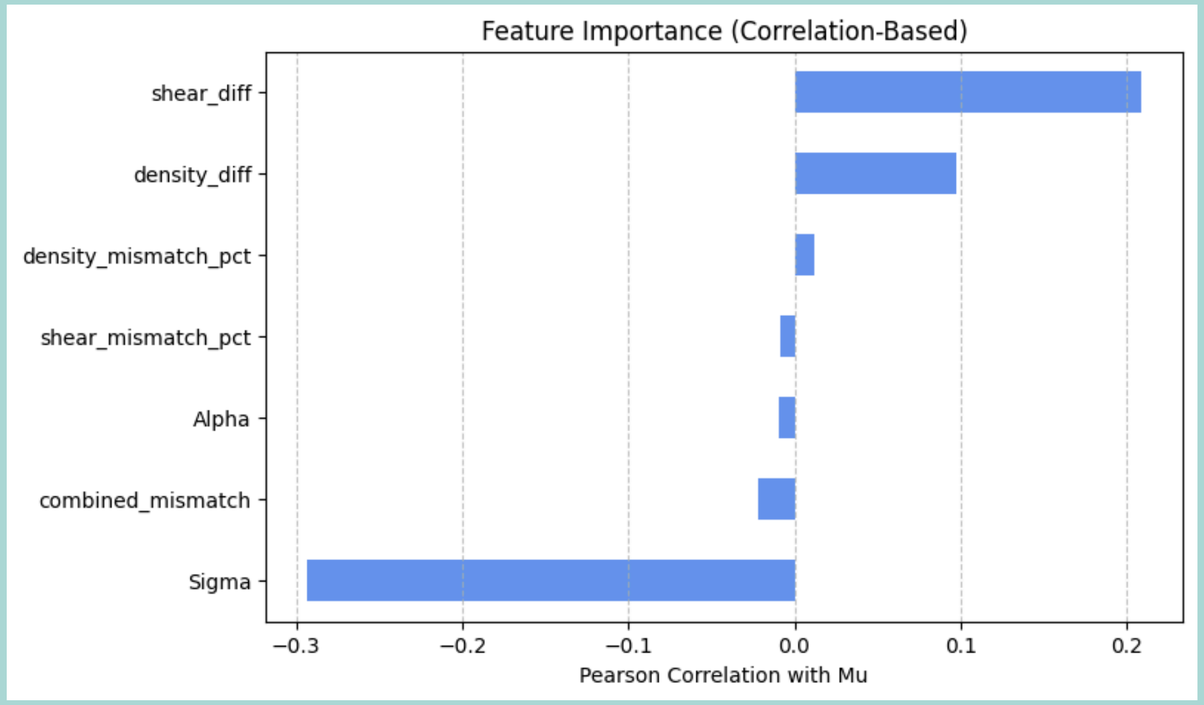
- **GINConv Layers:** These layers are effective at aggregating information from neighboring nodes and distinguishing different graph structures
- **Global Mean Pooling:** This aggregates the node-level features into a single graph-level representation
- **Fully Connected Regression Head:** This part of the network is responsible for mapping the learned graph features to the final continuous output
- **Optimizer:** The Adam optimizer is used to adjust the model's weights



Model Training and Validation Performance

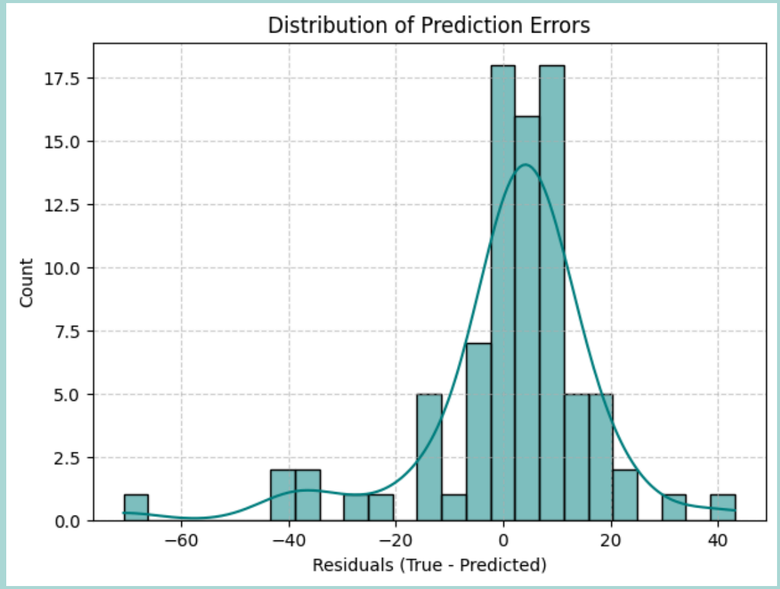
- During training, the GNN showed a steady decrease in Train MSE, indicating effective learning of solute–solvent interactions and segregation energy patterns
- Validation RMSE began to rise after several epochs, suggesting overfitting as the model started memorizing training data
- Early stopping was applied to maintain a balance between accuracy and generalization, ensuring the model captured meaningful, physics-based relationships across alloy systems

Feature Importance



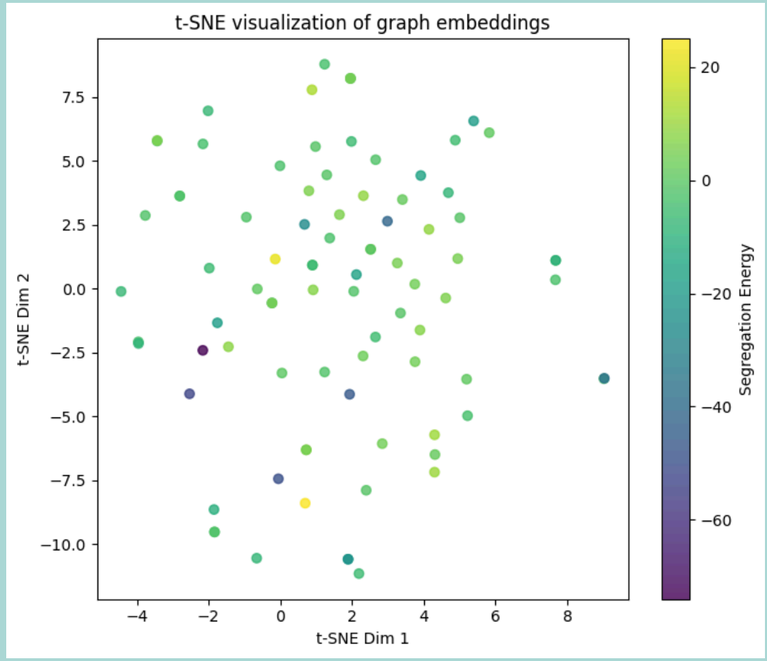
- shear_diff and density_diff show strongest positive correlations → mismatch in mechanical properties between solute and solvent significantly affects segregation energy
- Sigma has negative correlation → boundary character (structural mismatch) influences whether segregation is favorable/unfavorable
- Segregation energy is strongly tied to elastic and structural mismatch between elements at GBs.

Distribution of Prediction Errors



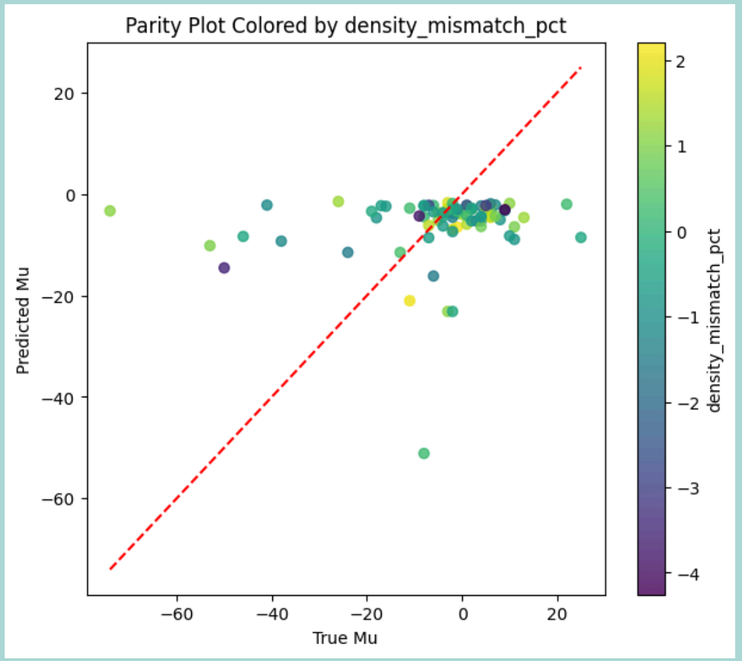
- The model performs reliably for most grain boundaries, with errors centered near zero. However, outliers indicate difficult-to-predict cases, suggesting the need for more diverse data or improved feature engineering to better capture complex GB behaviors.

t-SNE Visualization



- This confirms that the GNN has learned latent structure from atomic and thermophysical features that correlate with segregation energy

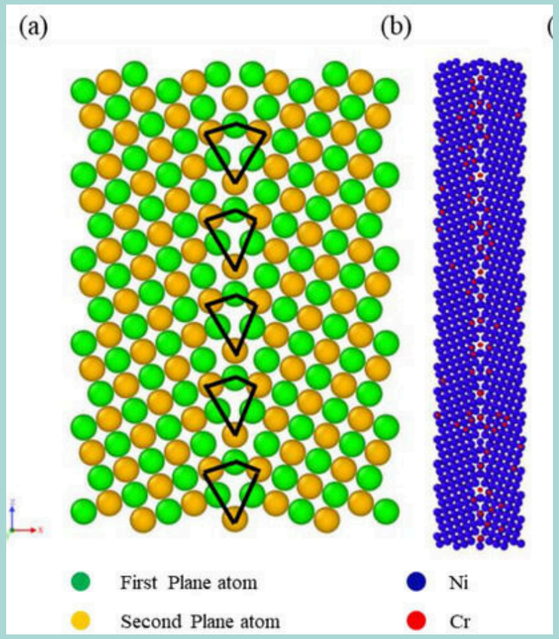
True vs Predicted

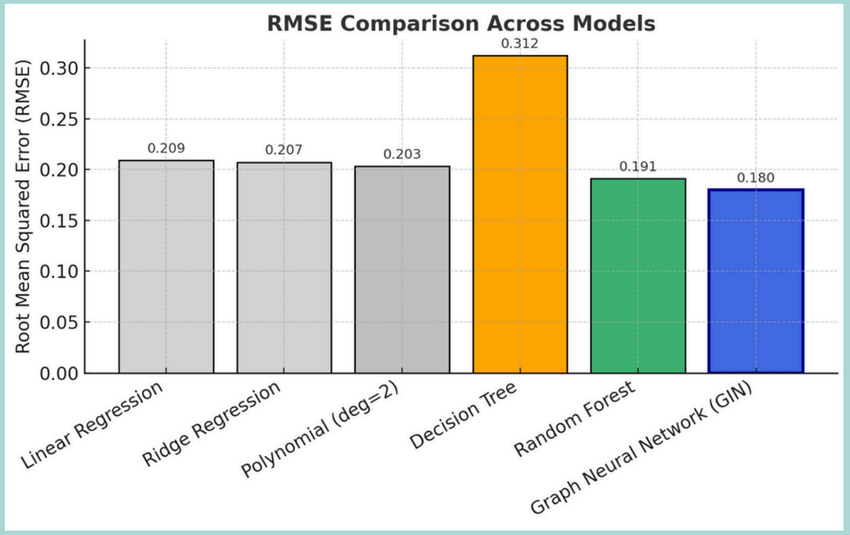


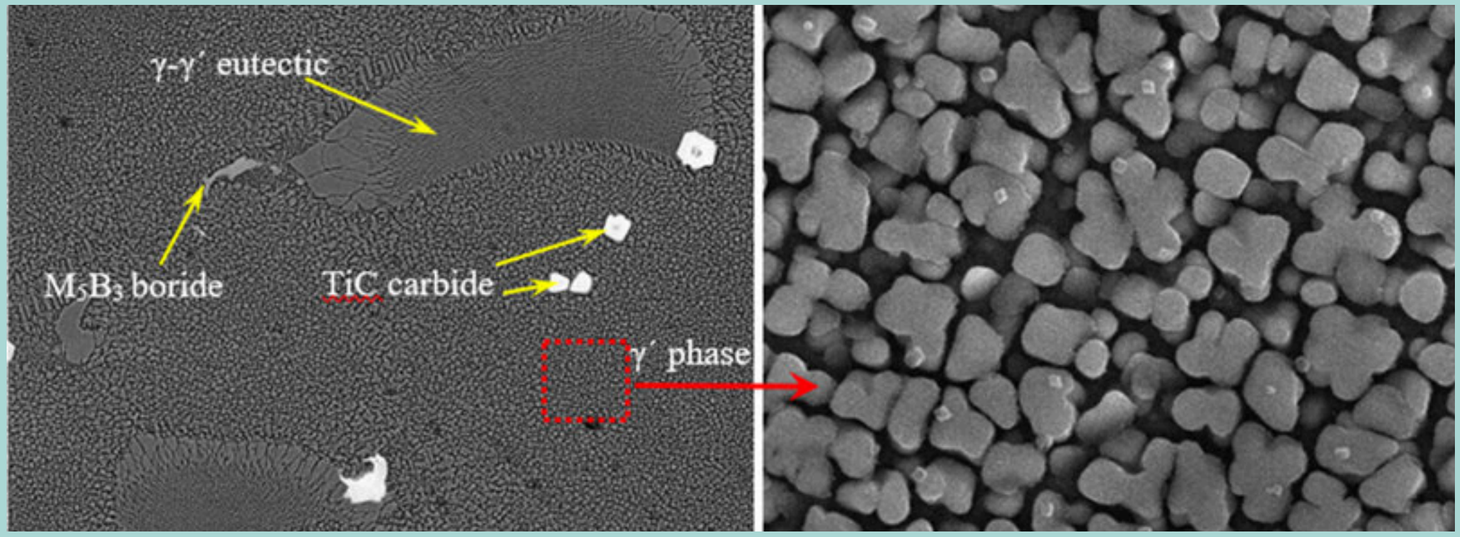
- Model predictions align well for moderate energy values but deviate for higher segregation energies, suggesting data sparsity and complex GB interactions affect generalization

CONCLUSION

Model	MSE	RMSE	R^2	Remarks
Linear Regression	0.0438	0.209	0.18	Captures basic trend only
Ridge Regression	0.0429	0.207	0.23	Slight regularization improvement
Polynomial (deg=2)	0.0415	0.203	0.38	Adds nonlinearity, limited gain
Decision Tree	0.0972	0.312	−0.610	Overfits; poor generalization
Random Forest	0.0367	0.191	0.54	Best among traditional models
Graph Neural Network (GIN)	**0.029–0.032 (train)0.034–0.038**	**0.170–0.195**	**≈0.68–0.72**	Best overall; learns atomic-level features







Thank you very much!