Title: Prediction Of Bandgap, Formation Energy and Bulk modulus of IIIA-VA group materials using ML model(XGBoost)

Sunitha Pavani 22CSB0B11
Rishmitha.R 22CSB0C15

राष्ट्रीय प्रौद्योगिकी संस्थान वारंगल
National Institute of Technology Warangal
An Institute of National Importance

Machine Learning for materials informatics (MLMI)

IIIA-VA Semiconductor Materials

- **III–V semiconductors are compounds formed by combining elements from**

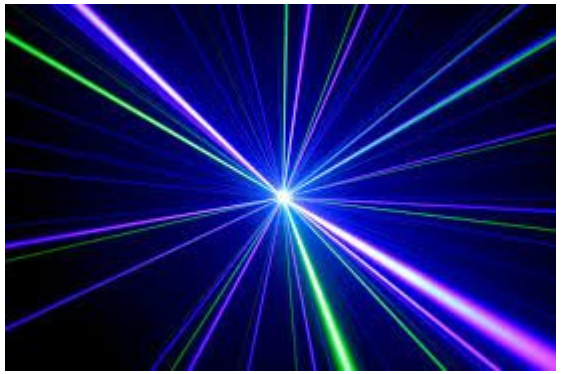
➢ ***Examples*** *:* GaAs, InP, AlN, GaN, InSb, AlP

IIIA		VA
B	C	N
Al	Si	P
Ga	Ge	As
In	Sn	Sb
Tl	Pb	Bi
Nh	Fl	Mc

- They are covalent crystals with partial ionic character.
- Similar crystal structures – most crystallize in zinc-blende or wurtzite lattices, enabling consistent property comparison.
- All are III–V semiconductors – exhibit direct or indirect band gaps

Applications Of IIIA-VA Materials

- Light-Emitting Diodes (LEDs)
- Laser Diodes
- High-Speed & RF Electronics
- Solar Cells
- Infrared Detectors & Sensors

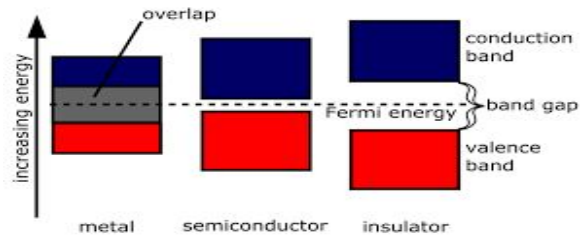


3

Properties to Predict

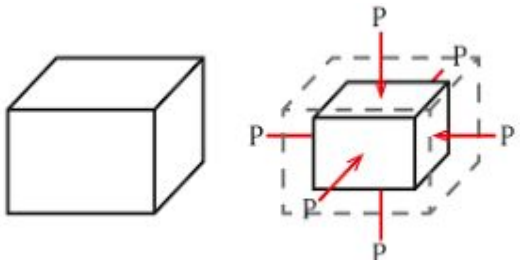
Bandgap:

- Energy difference between the valence band and conduction band.
- Controls optical absorption and electronic conductivity.



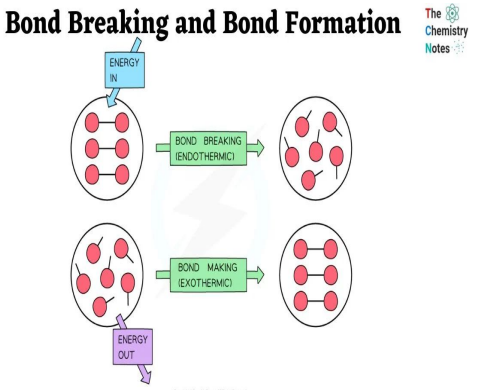
Bulk Modulus:

- The bulk modulus is a measure of a material's resistance to uniform compression
- Related to mechanical stiffness and bonding strength.



Formation Energy:

- Energy released or absorbed when a compound is formed from its constituent elements.
- Indicates thermodynamic stability of the material.



4

Model Overview – XGBoost

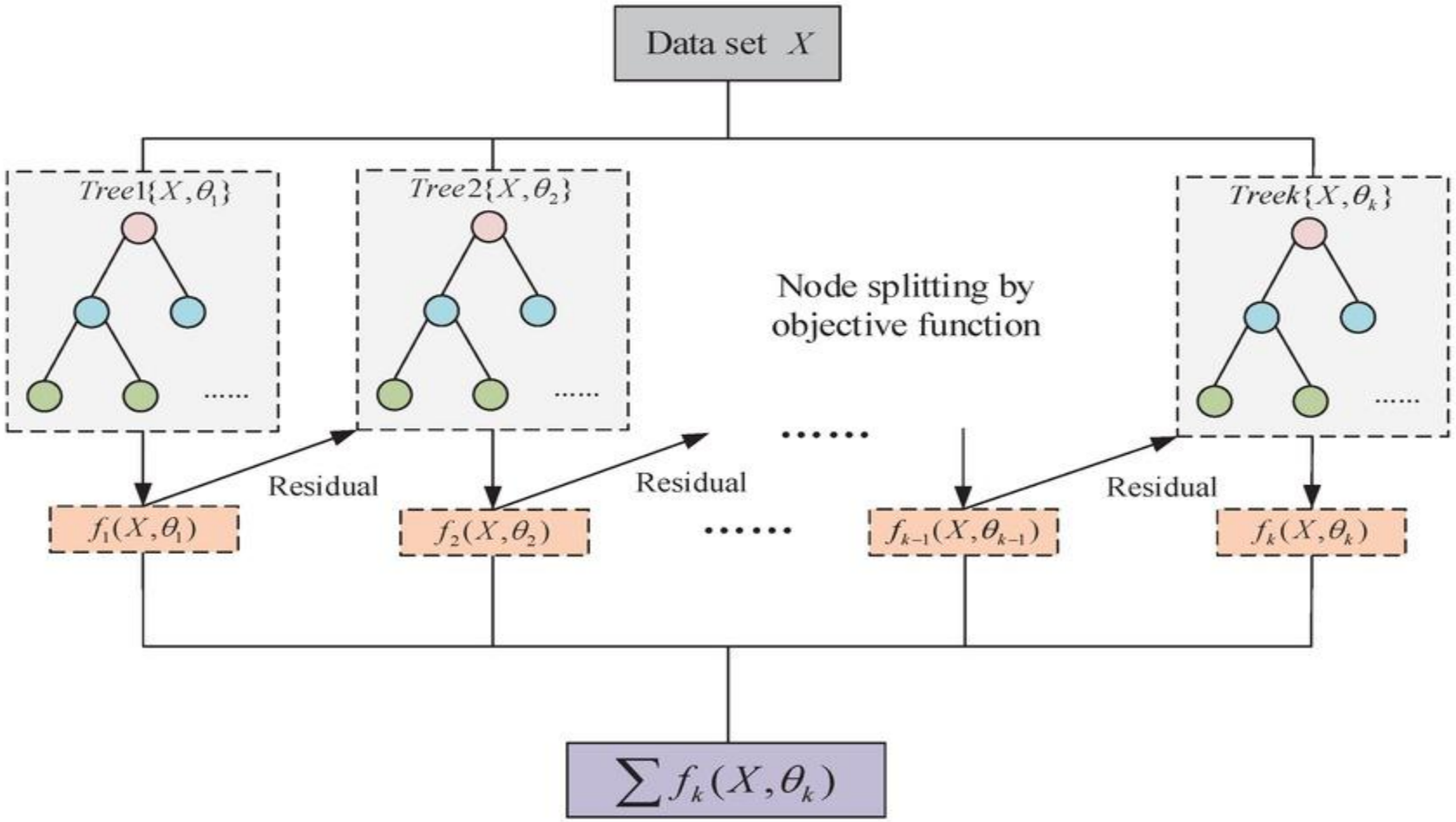
- XGBoost is a machine learning algorithm based on Gradient Boosted Decision Trees (GBDT).

- **Working Principle**:

➢ Starts with an initial prediction (e.g., mean of data).
➢ Builds a new decision tree to predict those residuals.
➢ Adds the new tree's prediction (scaled by learning rate) to improve results.
➢ Repeats this process until the error is minimized.

5

Model Overview – XGBoost



6

XGBoost Regression Model

- **XGBoost (Extreme Gradient Boosting): A high-performance ensemble learning method based on gradient-boosted decision trees.**

- **Gradient Boosting Update Rule:**
  $\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \eta * f_t(x_i)$
- **Objective Function:**
  $\mathcal{L}^{(t)} = \sum l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t)$
- **Regularization Term (for trees):**
  $\Omega(f) = \gamma * T + (1/2) * \lambda * \sum(w_j^2)$

- $\hat{y}_i^{(t)}$ → prediction for sample i at iteration t
- $f_t(x_i)$ → new tree (weak learner) output at iteration t
- η → learning rate
- l → loss function (e.g., MSE for regression)
- $\Omega(f_t)$ → regularization term for tree complexity
- T → number of leaves in the tree
- w_j → weight of leaf j
- γ, λ → regularization hyperparameters

Why XGBoost?

- **Learns complex nonlinear relationships between structure & properties**
- **Handles small scientific datasets effectively**
- **Built-in regularization to reduce overfitting**
- **Supports feature importance for model interpretability**
- **Highly optimized for speed, parallelism, and scalability**

Features and Labels

Label	Physical Domain	Features
Bandgap (optb88vdw_bandgap)	Electronic	is_orthogonal, mp_ratio, avg_valence, structural_density, atoms_per_cell, mass_diff, mp_diff, lattice_a, lattice_b, en_diff, r_sum_over_mass_sum, alpha, avg_lattice_constant, lattice_c, unit_cell_volume
Bulk Modulus (bulk_modulus_kv)	Mechanical / Elastic	Structural: a, b, c, alpha, beta, gamma, min_dist, max_dist, avg_dist Compositional: Z_mean, mass_mean, en_diff
Formation Energy (formation_energy_peratom)	Thermodynamic / Stability	mass_sum, structural_density, avg_valence, avg_molar_volume, lattice_b, Z_diff, avg_melting_point, polarizability_est, spg_number, Z_sum

Formulas for the calculation of predicting properties

Property	Formula	Parameters
Bandgap	Bandgap (E_g): $E_g = E\_C - E\_V$	EC = energy at conduction band minimum EV = energy at valence band maximum
Bulk Modulus	Bulk Modulus (B): $B = -V * (dP/dV)$ Alternative (from energy–volume curve): $B = V * (d^2E/dV^2)$	V = volume P = pressure E = total energy of the material
Formation Energy	Formation Energy (E_f): $E\_f = [E\_total(A_xB_Y) - x*\mu\_A - y*\mu\_B] / (x + y)$	E_total(AxBy) = total DFT energy of compound AxBy μA = chemical potential of element A μB = chemical potential of element B x = number of atoms of A y = number of atoms of B

Loss Function

Training Loss Function: Mean Squared Error (MSE)

Mean Squared Error (MSE): $MSE = (1/n) \times \Sigma (y_i - \hat{y}_i)^2$

Parameters:

n = total number of data points

y_i = actual (true) value

$\hat{y}_i$ = predicted value

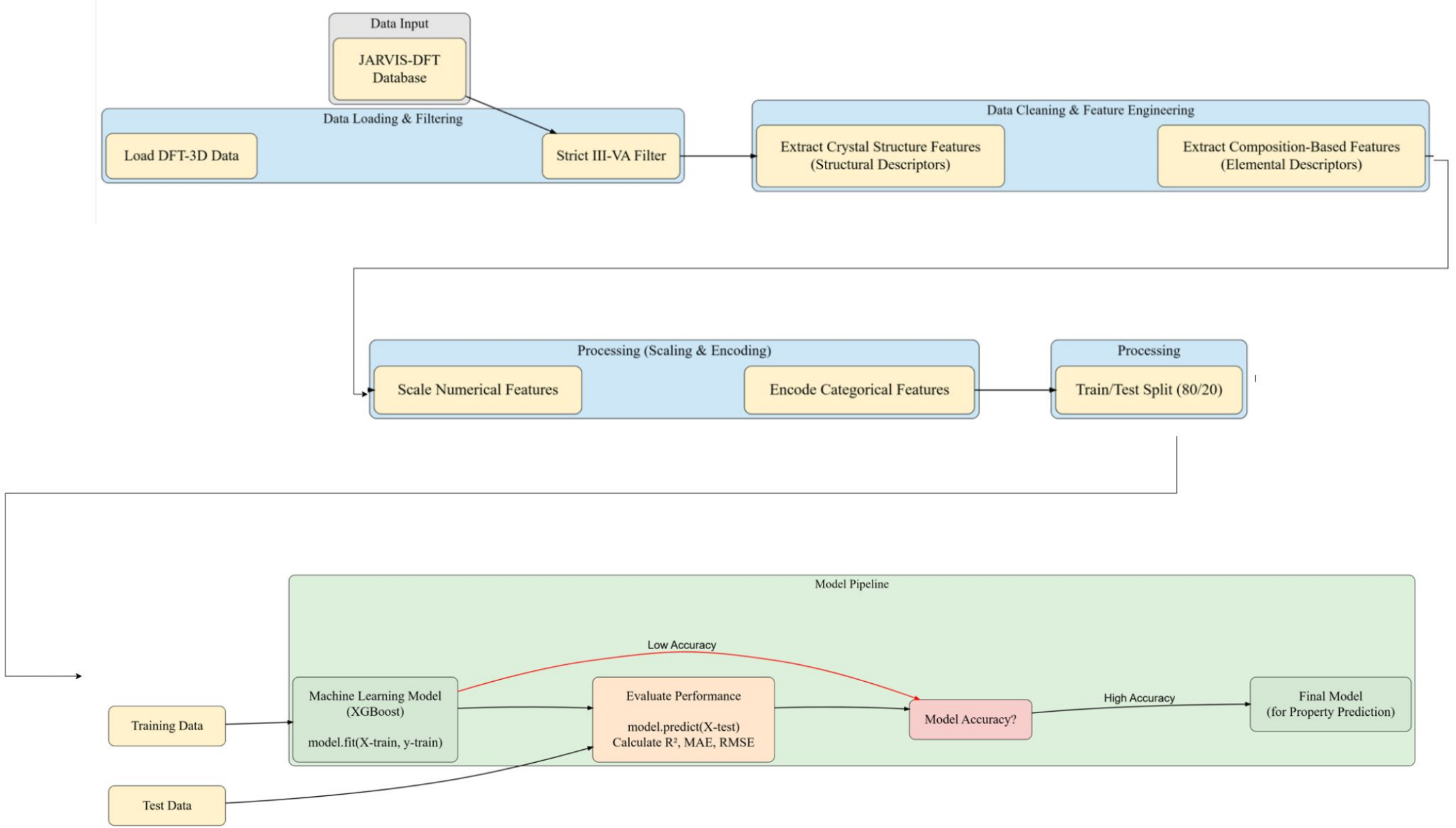
Why MSE?

- Penalizes larger errors more heavily due to squaring.
- Encourages predictions to be close to true values on average.
- Smooth gradient for optimization → better convergence in tree-based models.

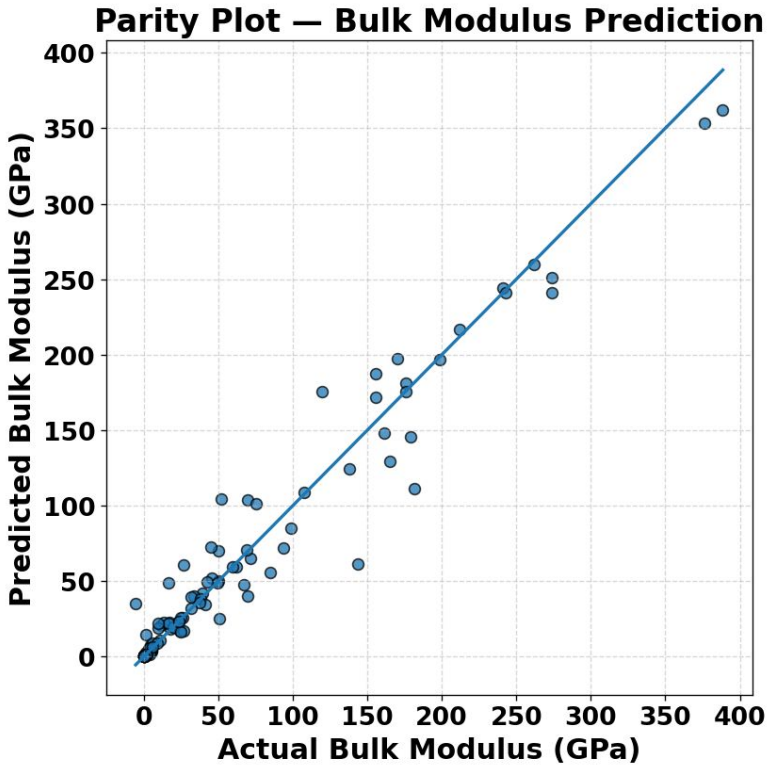
Evaluation Metrics

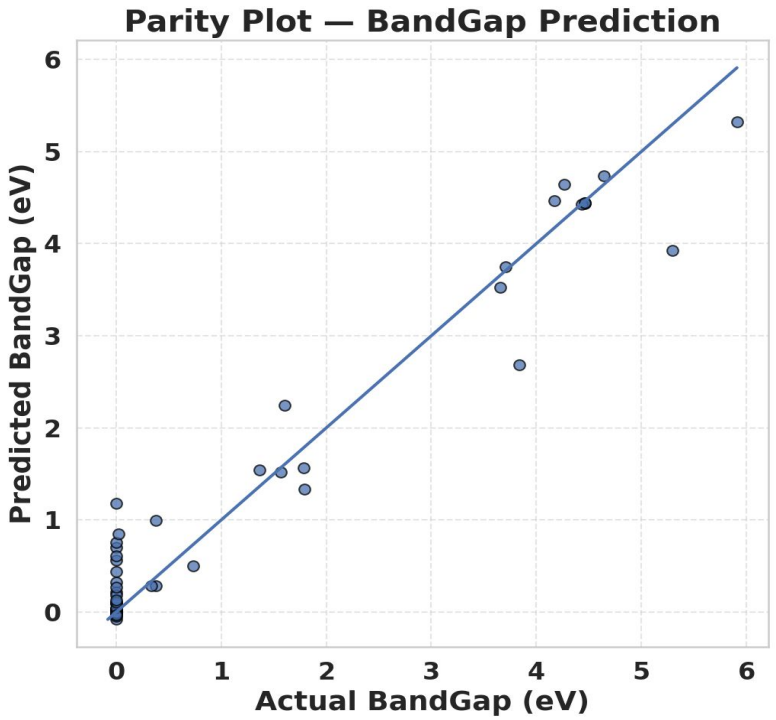
- R^2 (Coefficient of Determination): $R^2 = 1 - [\ \Sigma (y_i - \hat{y}_i)^2 / \Sigma (y_i - \bar{y})^2\ ]$
- Mean Absolute Error (MAE): $MAE = (1/n) \times \Sigma\ |y_i - \hat{y}_i|$
- Root Mean Squared Error (RMSE): $RMSE = \sqrt{[\ (1/n) \times \Sigma (y_i - \hat{y}_i)^2\ ]}$

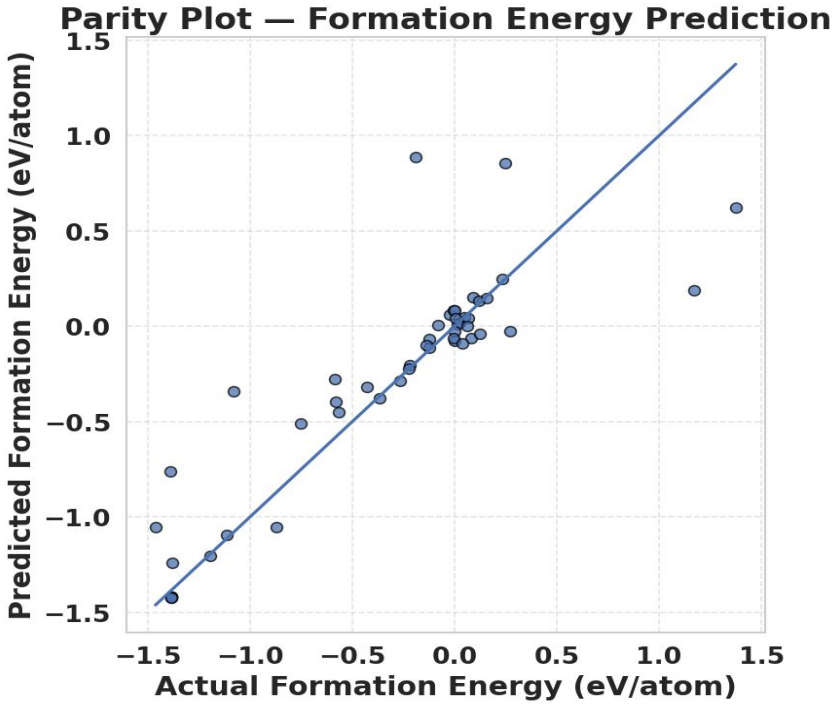
Pipeline for Predicting Target Properties



Parity Plots & Accuracy







	Train R2	Test R2	MAE	RMSE
Bandgap (eV)	0.994	0.949	0.2689	0.4362
Bulk Modulus (GPa)	0.999	0.946	16.971	26.446
Formation Energy (eV/atom)	0.996	0.736	0.153	0.334

Conclusions

- To overcome the high computational cost and time required by DFT for screening new IIIA–VA materials, a machine-learning-based predictive framework using XGBoost has been developed.

- We constructed ML models to estimate Band Gap, Formation Energy, and Bulk Modulus using reliable JARVIS-DFT data and carefully engineered structural and compositional features.

- The predicted results show strong agreement with DFT-reported values, demonstrating that the proposed ML approach can efficiently capture key material property trends while drastically reducing computation time.

- This framework provides a promising pathway for rapid materials discovery and optimization in semiconductor applications, enabling large-scale screening of unexplored III–V compounds.