Machine Learning for Rapid Prediction of Exfoliation Energies in 2D Materials

Artemiy Filippov*

Department of Computational Mathematics, Science and Engineering

Michigan State University, East Lansing, MI 48824

(Dated: December 7, 2025)

Abstract

Two-dimensional (2D) materials such as graphene and transition-metal dichalcogenides possess exceptional electronic and mechanical properties, but discovering new candidates requires computing their exfoliation energies, the energy needed to separate a monolayer from its bulk structure. These energies are typically obtained through Density Functional Theory (DFT), which is accurate but computationally expensive and limits large-scale screening efforts. In this work, I develop machine learning models to rapidly predict exfoliation energies using the MatBench JDFT-2D dataset (635 materials) and a set of compositional and structural descriptors generated with Matminer. After feature selection reduced the descriptor set from 136 to 92 features, three models were evaluated: Linear Regression, Random Forest, and a shallow Neural Network. Random Forest achieved the best performance with a test MAE of 31.8 meV/atom and $R^2 \approx 0.50$, representing a 45% improvement over the linear baseline. SHAP analysis reveals that the model relies on physically meaningful features such as volume per atom and Mendeleev-number statistics. All models, however, systematically underpredict exfoliation energies above 300 meV/atom due to an extreme 20:1 class imbalance. Overall, this study demonstrates that machine learning provides an efficient tool for large-scale screening of layered materials, while highlighting that improved datasets or physics-informed models are needed to achieve reliable performance in the high-energy regime.

BACKGROUND AND MOTIVATION

Since the isolation of graphene in 2004 [1], two-dimensional (2D) materials have emerged as a frontier in materials science with applications spanning nanoelectronics, energy storage, and quantum computing [2]. Exfoliation energy (E_{exf}) determines whether a layered material can be isolated as a stable monolayer, with low energies (below 200 meV/atom) enabling mechanical or liquid-phase separation [3].

However, calculating E_{exf} using Density Functional Theory (DFT) requires 20-100 CPU hours per material with van der Waals corrections and stringent convergence criteria. As materials databases grow to hundreds of thousands of compounds, DFT screening becomes prohibitively costly for high-throughput discovery workflows.

Machine learning provides a practical alternative by learning structure-property relationships from existing DFT data. Once trained, ML models predict exfoliation energies from compositional and structural descriptors in milliseconds, enabling rapid pre-screening. Prior work demonstrates ML can achieve DFT-level accuracy for formation energies and bandgaps [5], motivating application to exfoliation prediction.

This project develops and compares three ML models—Linear Regression, Random Forest, and Neural Networks—to predict exfoliation energies from 635 2D materials, enabling high-throughput discovery while identifying limitations for extreme material properties.

DATA DESCRIPTION

Data Origins

I use the **MatBench JDFT-2D** dataset [6], a curated benchmark from the Materials Project database [4]. The dataset consists of DFT-computed exfoliation energies for experimentally known and hypothetical 2D materials, calculated using the Perdew–Burke–Ernzerhof (PBE) generalized gradient approximation (GGA) exchange–correlation functional, supplemented with the optB88-vdW van der Waals correction and implemented in VASP [7]. Each material's exfoliation energy was computed as:

$$E_{\text{exf}} = \frac{E_{\text{mono}} - E_{\text{bulk}}/n}{N_{\text{atoms}}} \tag{1}$$

where E_{mono} is the total energy of an isolated monolayer, E_{bulk} is the bulk crystal energy, n is the number of layers, and N_{atoms} is the number of atoms per layer. All calculations were performed with consistent convergence criteria (energy convergence < 0.001 eV, force convergence < 0.01 eV/Å).

Dataset Characteristics

The dataset contains 635 layered compounds with DFT-computed exfoliation energies. Each material includes a full crystal structure, and compositional and structural descriptors were generated using Matminer. The target is exfoliation energy measured in meV/atom.

Data Quality Analysis

Target Distribution and Class Imbalance

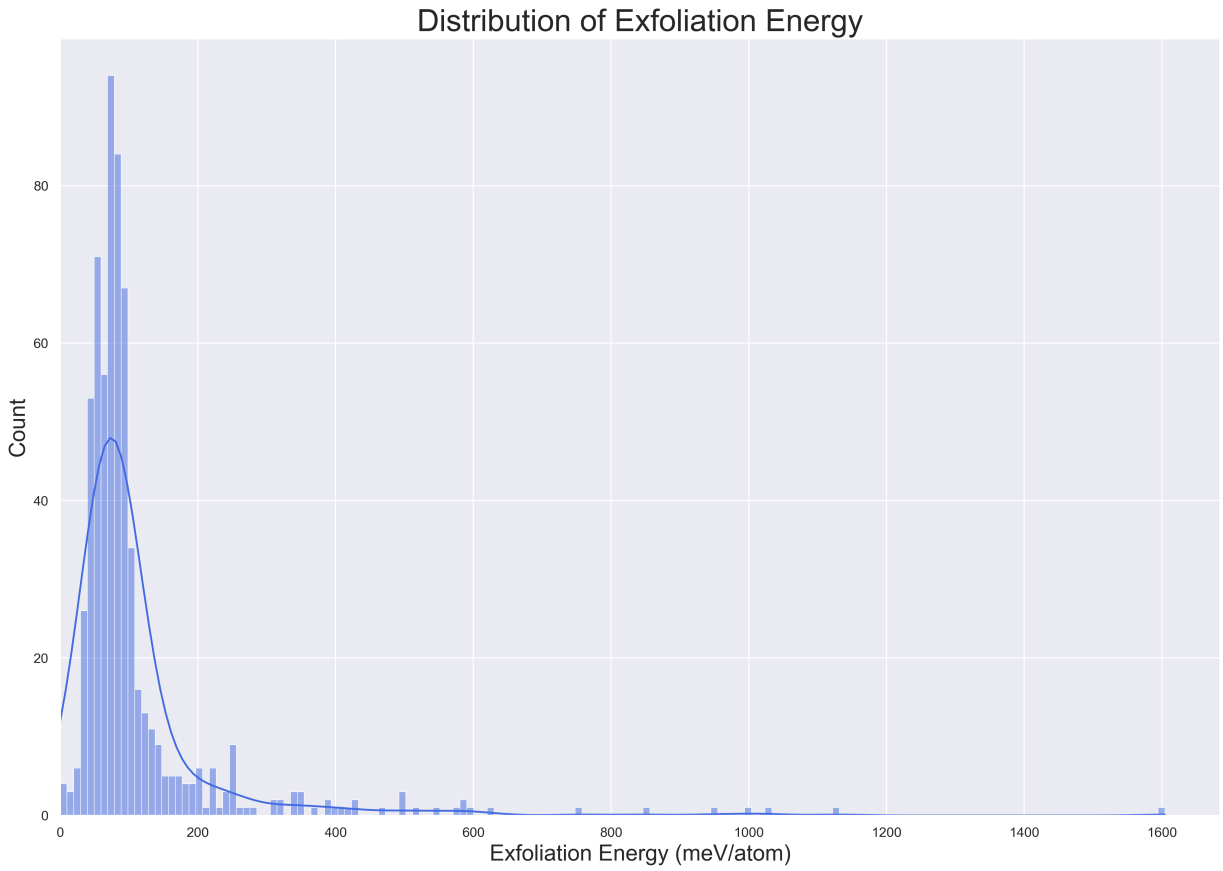

FIG. 1: Distribution of exfoliation energies in the MatBench JDFT-2D dataset. The histogram shows strong right-skewness with most materials below 100 meV/atom, reflecting the prevalence of weakly-bonded layered structures like graphite and TMDs.

Feature Correlation Analysis

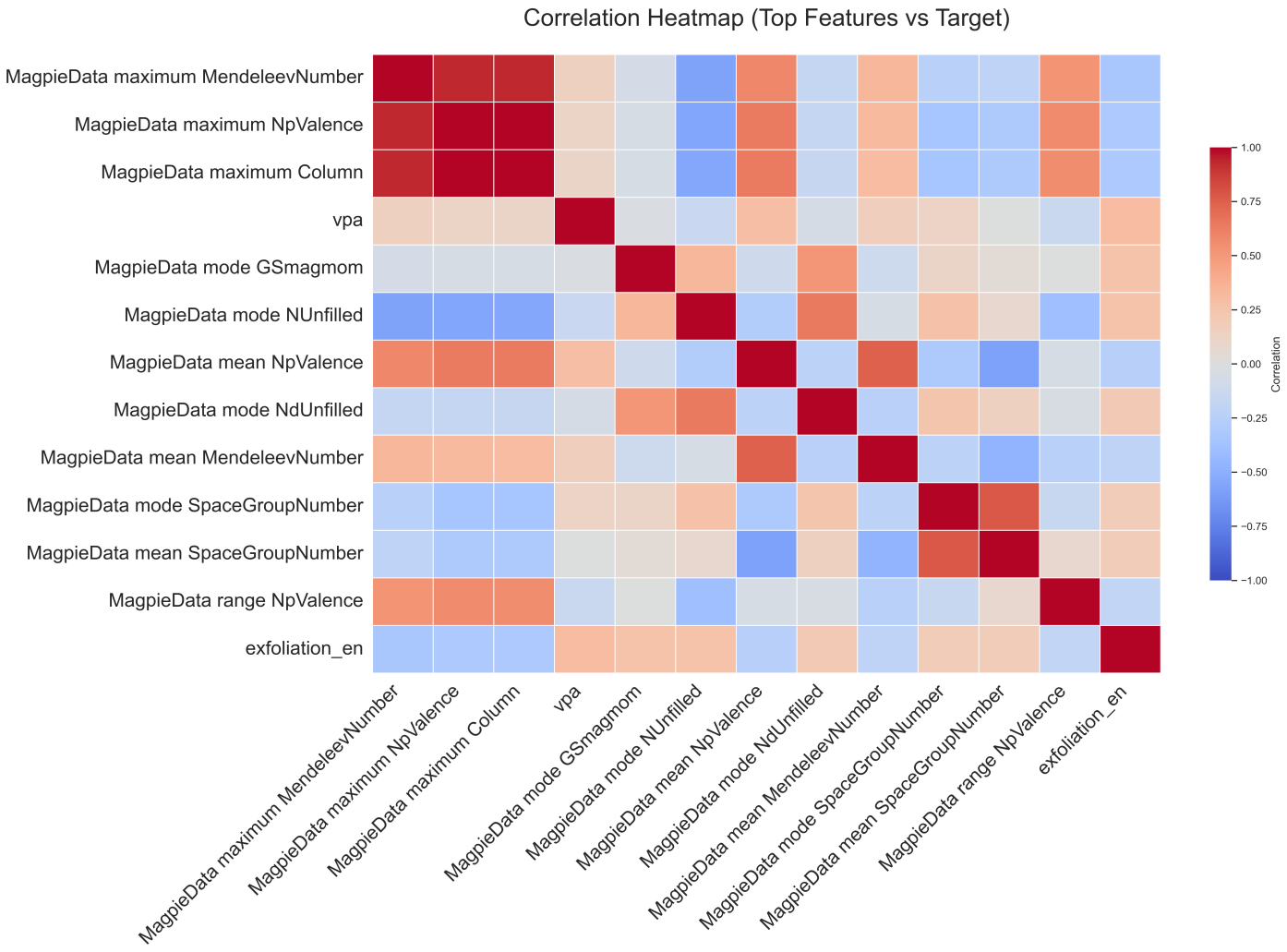

FIG. 2: Correlation heatmap of top 12 features versus exfoliation energy. Moderate correlations ($|r| < 0.35$) indicate that no single descriptor fully captures exfoliation behavior, motivating the use of ensemble machine learning methods.

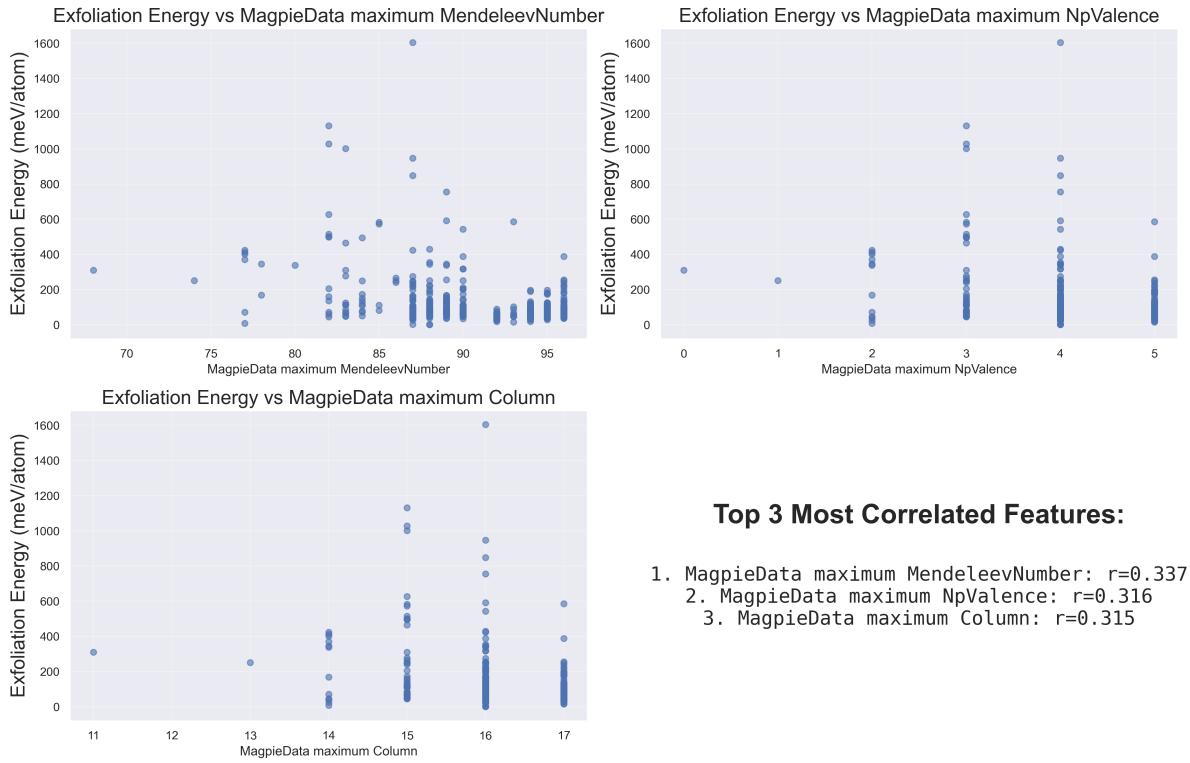

FIG. 3: Scatter plots of exfoliation energy versus top three correlated features. Non-linear patterns and heteroscedasticity motivate the use of flexible, non-parametric models like Random Forest.

PREPROCESSING

Feature Generation Pipeline

Raw crystal structures were converted to 135 numerical descriptors using Matminer's featurization framework [8], including density features (volume per atom, packing fraction) and Magpie compositional statistics. A two-stage feature selection reduced dimensionality to 92 features by removing low-variance (< 0.01) and highly correlated ($|r| > 0.95$) descriptors (Figure 4), retaining all top-10 correlated features (Table I). An 80/20 train-test split yielded 509 training and 127 test materials, with StandardScaler normalization applied uniformly.

Feature Selection

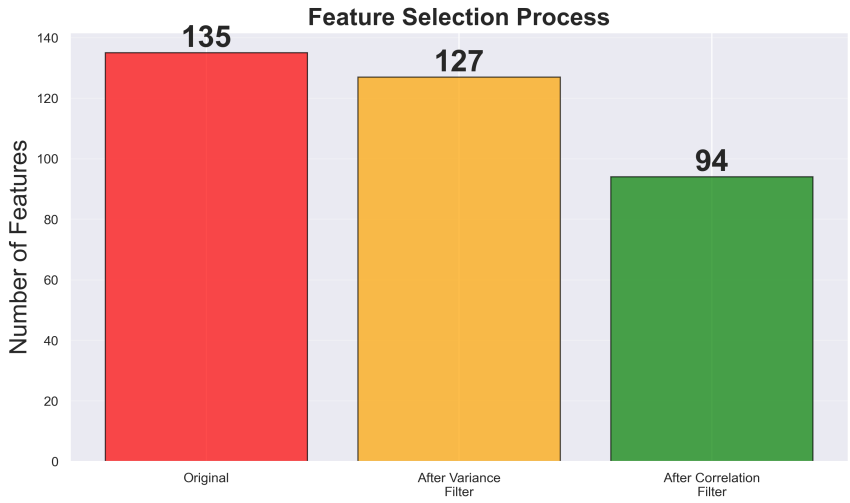

FIG. 4: Feature selection pipeline reduced dimensionality from 135 to 92 features by removing low-variance and highly correlated descriptors, improving model generalization.

TABLE I: Top 10 features by absolute correlation with exfoliation energy (all retained after feature selection).

Rank	**Feature**	$\|r\|$
1	MagpieData maximum MendeleevNumber	0.3323
2	vpa (volume per atom)	0.3124
3	MagpieData maximum NpValence	0.3119
4	MagpieData maximum Column	0.3106
5	MagpieData mode GSmagmom	0.2624
6	MagpieData mode NUnfilled	0.2604
7	MagpieData mean NpValence	0.2465
8	MagpieData mode NdUnfilled	0.2136
9	MagpieData mean MendeleevNumber	0.2080
10	MagpieData mean SpaceGroupNumber	0.1893

Data Splitting

I employed an 80/20 stratified train-test split with `random_state=42` for reproducibility, yielding 509 training materials and 127 test materials, with five-fold cross-validation performed exclusively on the training set to avoid leakage.

MACHINE LEARNING TASK AND OBJECTIVE

Why Machine Learning?

DFT methods pose barriers in computational cost, required expertise, and scalability. Each exfoliation calculation demands tens of CPU hours with iterative self-consistent field cycles and careful numerical settings. The workflow also requires expert knowledge of exchange–correlation functionals and convergence criteria. As a result, screening even modestly sized candidate sets becomes infeasible; evaluating ten thousand materials would require on the order of two million CPU hours.

Task Type

The problem is a supervised regression task in which exfoliation energies are continuous targets derived from paired feature–label data. Because test compounds occupy the same chemical space as the training distribution, the task is primarily interpolative rather than extrapolative.

MODELS

Three models of increasing complexity were compared: Linear Regression (Ridge, $\lambda = 0.01$) as baseline, Random Forest (500 trees, max depth 20) for nonlinear interactions, and a shallow Neural Network ($92 \to 128 \to 64 \to 1$ with L_2 regularization $\alpha = 0.01$). See Table II for full specifications.

TRAINING METHODOLOGY

Loss Functions

Linear Regression:

$$\mathcal{L}(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mathbf{w}^T\mathbf{x}_i)^2 + \lambda||\mathbf{w}||_2^2 \tag{2}$$

Random Forest:

Each tree minimizes mean squared error at splits:

$$\text{MSE} = \frac{1}{n_{\text{node}}}\sum_{i\in\text{node}}(y_i - \bar{y}_{\text{node}})^2 \tag{3}$$

Final prediction is the average over 500 trees.

Neural Network:

$$\mathcal{L}(\mathbf{W}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - f_{\mathbf{W}}(\mathbf{x}_i))^2 + \alpha\sum_{\ell}||\mathbf{W}_\ell||_2^2 \tag{4}$$

where $f_{\mathbf{W}}$ is the network output and $\mathbf{W}_\ell$ are layer weights.

Training Process

Overfitting Prevention

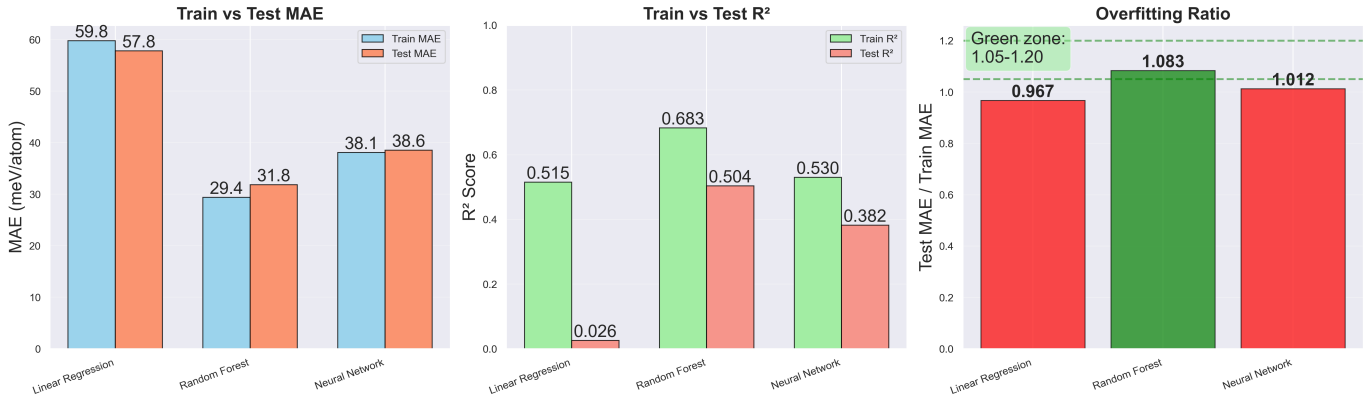

FIG. 5: Overfitting analysis for all models. All three models achieve test/train ratios around 0.96-1.08, indicating proper generalization without severe overfitting or underfitting. Random Forest shows the best overall performance with $\mathbf{R^2} = 0.504$ on test data.

Model Summary Table

TABLE II: Summary of models, parameters, and training configuration.

Model	Parameters	Hyperparameters	Loss	Regularization
Linear Regression	93 weights	No tuning	MSE	L2 Ridge
Random Forest	500 trees (∼1200 nodes/tree)	Max depth = 20; min split = 10; min leaf = 3	MSE per tree	Tree structural constraints
Neural Network	∼13,000 params (92→128→64→1)	Learning rate = 0.002; L_2 weight = 0.01; batch size = 64	MSE	L2 penalty + early stopping

METRICS

Primary Metric: Mean Absolute Error (MAE)

MAE measures average prediction error in the same units as the target:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{5}$$

MAE is directly interpretable in the physical units of the target, is less sensitive to outliers than MSE, and provides a clear statement of typical prediction error in meV/atom.

Secondary Metrics

Root Mean Squared Error (RMSE):

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{6}$$

Penalizes large errors more heavily than MAE.

Coefficient of Determination (R^2):

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \tag{7}$$

Fraction of variance explained. $R^2 = 1$ is perfect, $R^2 = 0$ is no better than predicting the mean.

Cross-Validation

Evaluation relies on five-fold cross-validation performed strictly within the training set, together with a final assessment on a held-out test set reserved for unbiased performance measurement.

RESULTS AND MODEL COMPARISON

Performance Comparison

TABLE III: Model performance on held-out test set (127 materials). Random Forest outperforms alternatives by 45% in MAE reduction vs. baseline.

Model	**Test MAE (meV/atom)**	**Test RMSE (meV/atom)**	**Test R^2**	**vs. Baseline (% improvement)**
Linear Regression	57.85	87.67	0.026	— (baseline)
Random Forest	**31.84**	**62.58**	**0.50**	**+44.95%**
Neural Network	38.55	69.80	0.382	+33.36%

Random Forest provides the best overall performance, achieving an MAE of 31.84 meV/atom. In contrast, linear regression fails to capture the strong nonlinear relationships in the data, as reflected by its low R^2. The neural network performs better than the linear model but still substantially lags behind Random Forest, likely due to the limited dataset size and the fact that Random Forest more naturally captures the threshold-type interactions present in the descriptors.

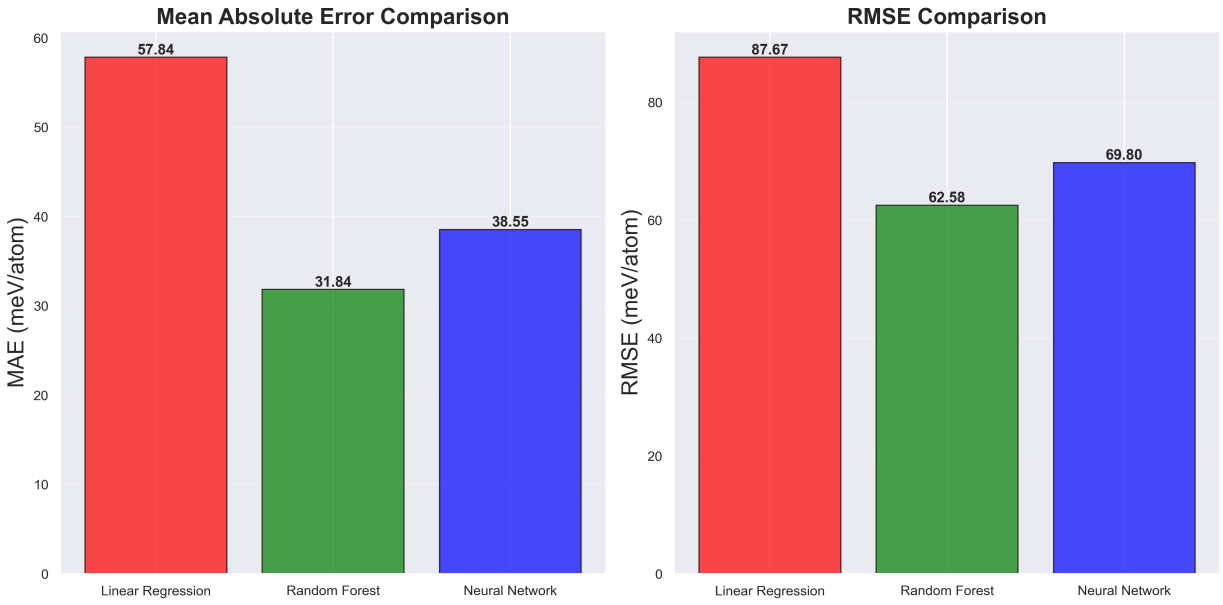

FIG. 6: Test set MAE and RMSE comparison. Random Forest (green) significantly outperforms Linear Regression (red) and Neural Network (blue), validating the importance of non-linear modeling for materials property prediction.

Prediction Quality Visualization

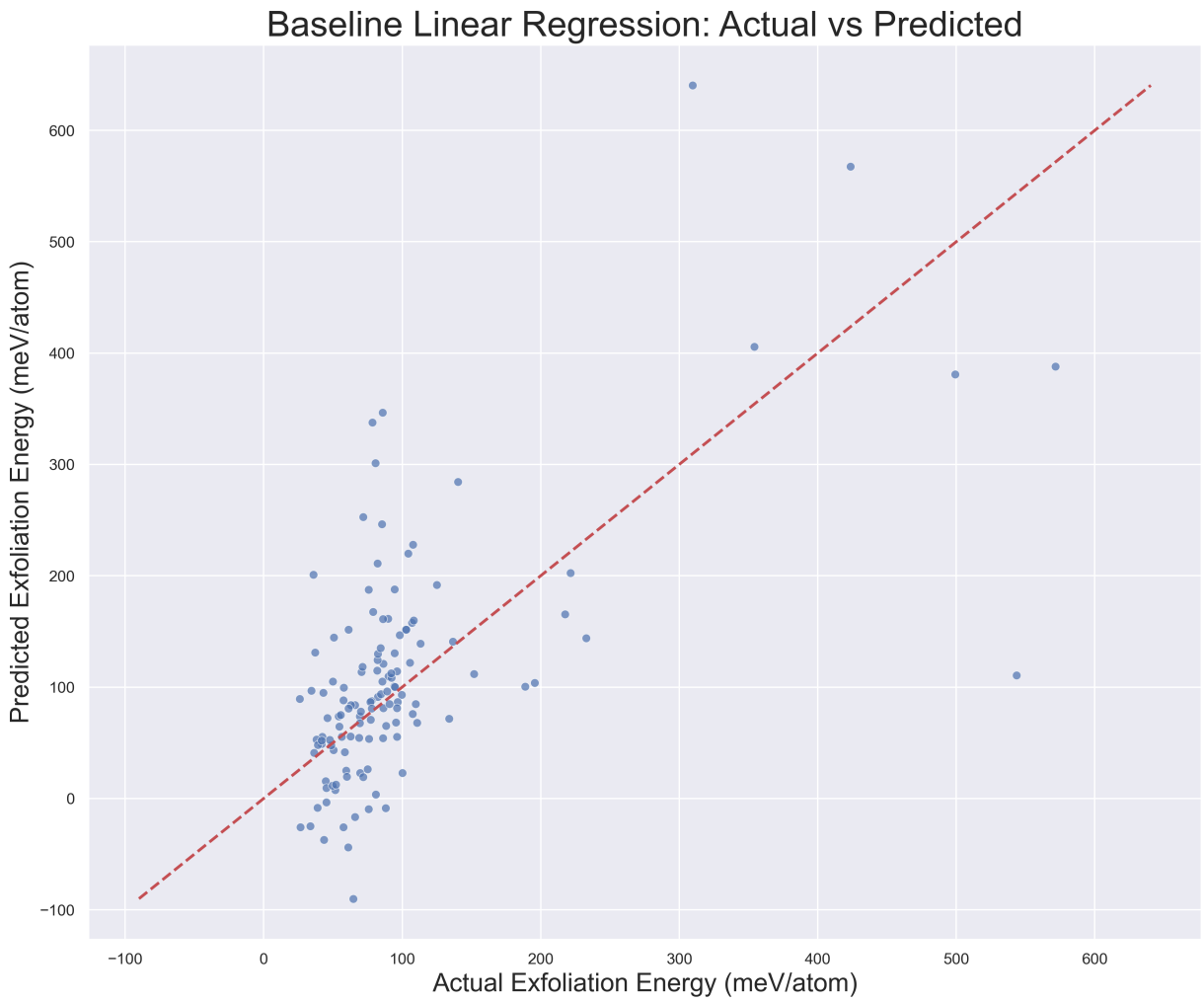

FIG. 7: Linear Regression predictions. Systematic underprediction of high-energy materials and large scatter confirm inadequacy of linear models for this task.

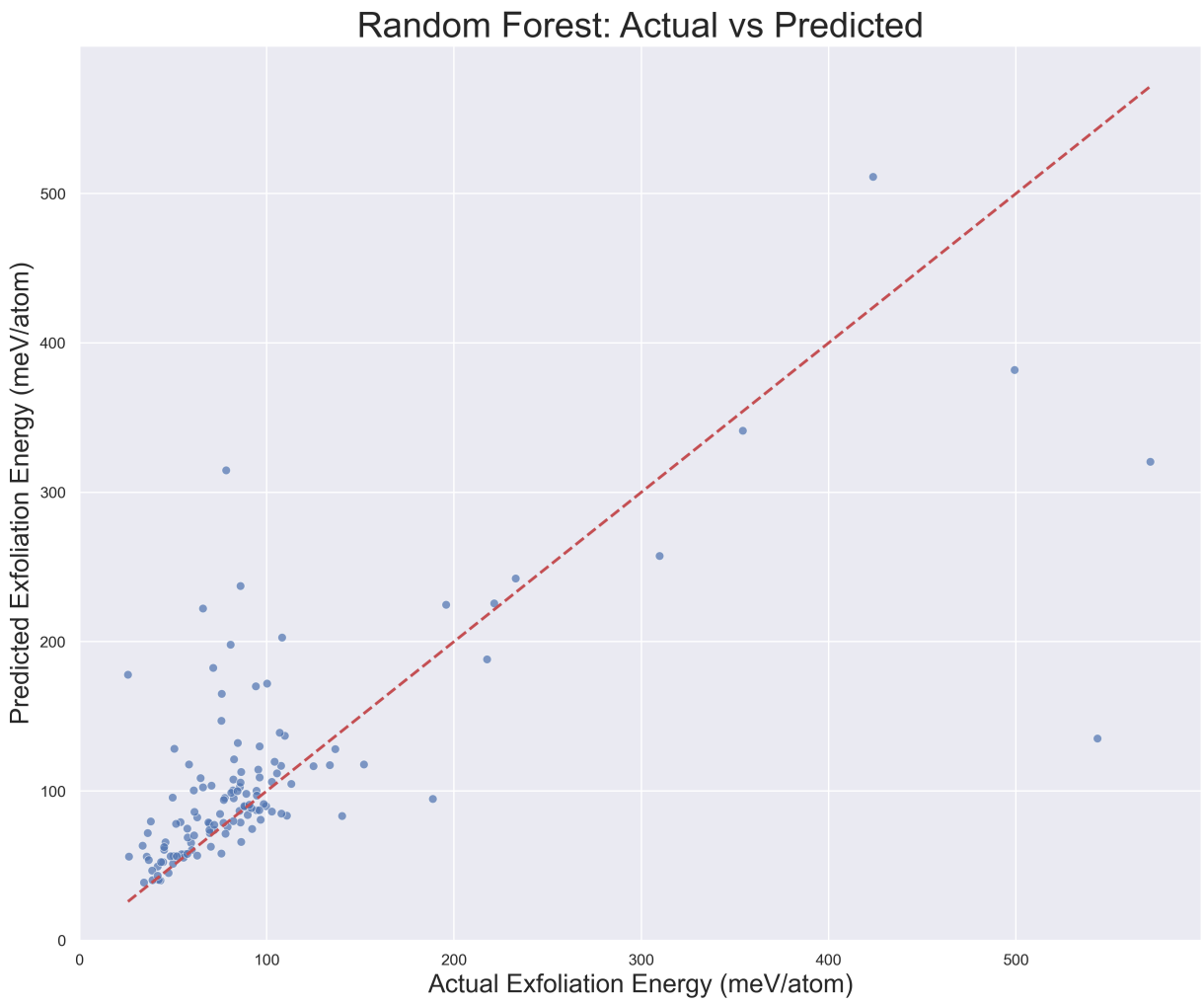

FIG. 8: Random Forest predictions. Tight clustering around the diagonal (red dashed line) for $E_{\text{exf}} < 300$ meV demonstrates strong predictive performance in the data-rich regime.

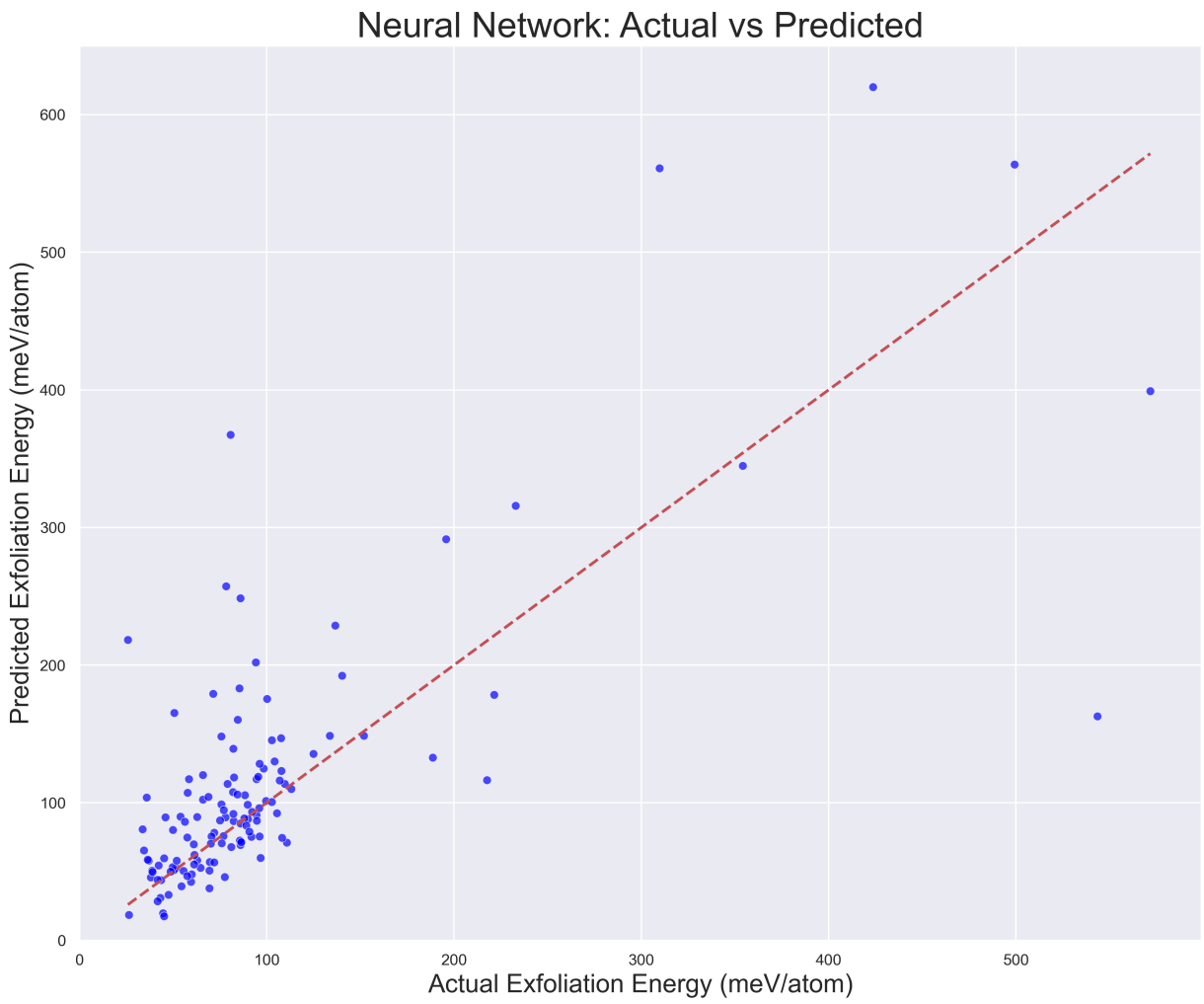

FIG. 9: Neural Network predictions. Performance intermediate between Linear Regression and Random Forest, with increased scatter at high energies indicating generalization challenges.

Computational Efficiency

TABLE IV: Training and inference times (Intel Core i7, 16GB RAM, no GPU). ML models achieve 4-6 orders of magnitude speedup vs. DFT.

Method	**Training Time**	**Inference Time (per material)**	**Speedup vs. DFT**
DFT (VASP)	N/A	20-100 CPU hours	1× (baseline)
Linear Regression	0.2 s	0.05 ms	$\sim 10^6\times$
Random Forest	12 s	2 ms	$\sim 10^5\times$
Neural Network	45 s	0.1 ms	$\sim 10^6\times$

Analysis and Discussion

Random Forest performs best because it offers non-parametric flexibility, naturally captures feature interactions, and reduces variance through ensemble averaging. These characteristics align well with structured tabular data [9–11], enabling the model to learn nonlinear relationships governing exfoliation energy.

Linear regression underfits the data, as evidenced by its train-test ratio of 0.97 (test error lower than train error), indicating the model is too simple to capture the nonlinear structure-property relationships.

The neural network underperforms primarily due to data scarcity: with only 635 samples, the model lacks the volume typically required for stable deep-learning optimization.

MODEL INTERPRETATION

Feature Importance via SHAP

I used SHAP (SHapley Additive exPlanations) [12] to interpret Random Forest's predictions. SHAP values quantify each feature's contribution to individual predictions, providing

both global importance rankings and local explanations.

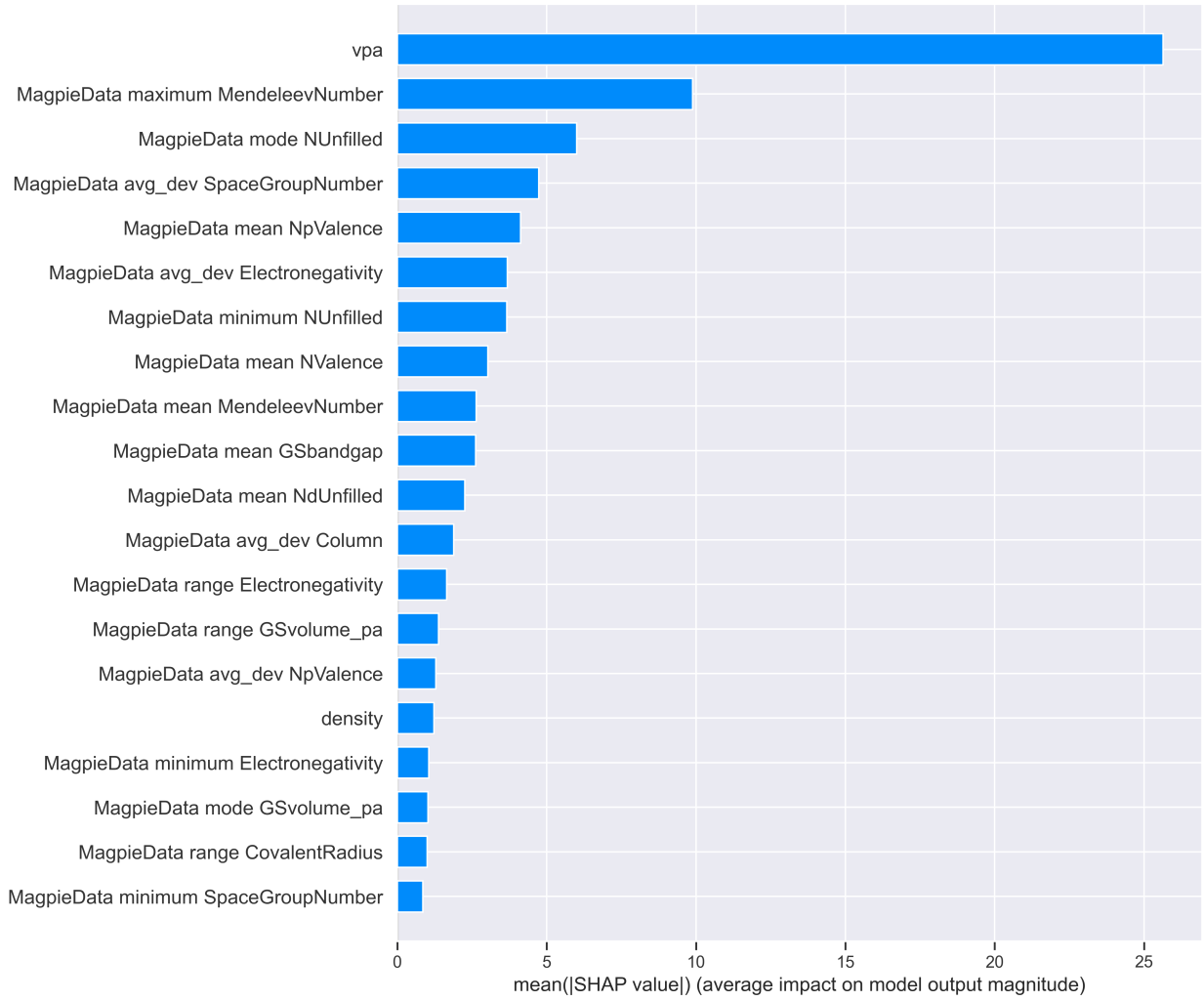

FIG. 10: SHAP feature importance (mean absolute SHAP value). Volume per atom (vpa) and compositional descriptors dominate, consistent with the physical intuition that interlayer spacing and element identity govern bonding strength.

The most influential features include volume per atom, which decreases predicted exfoliation energy as interlayer spacing grows; the mean and range of the bandgap, which relate to electronic structure and bonding strength; the average number of valence electrons; and the maximum Mendeleev number, which highlights the effect of heavier elements that typically form weaker van-der-Waals layers.

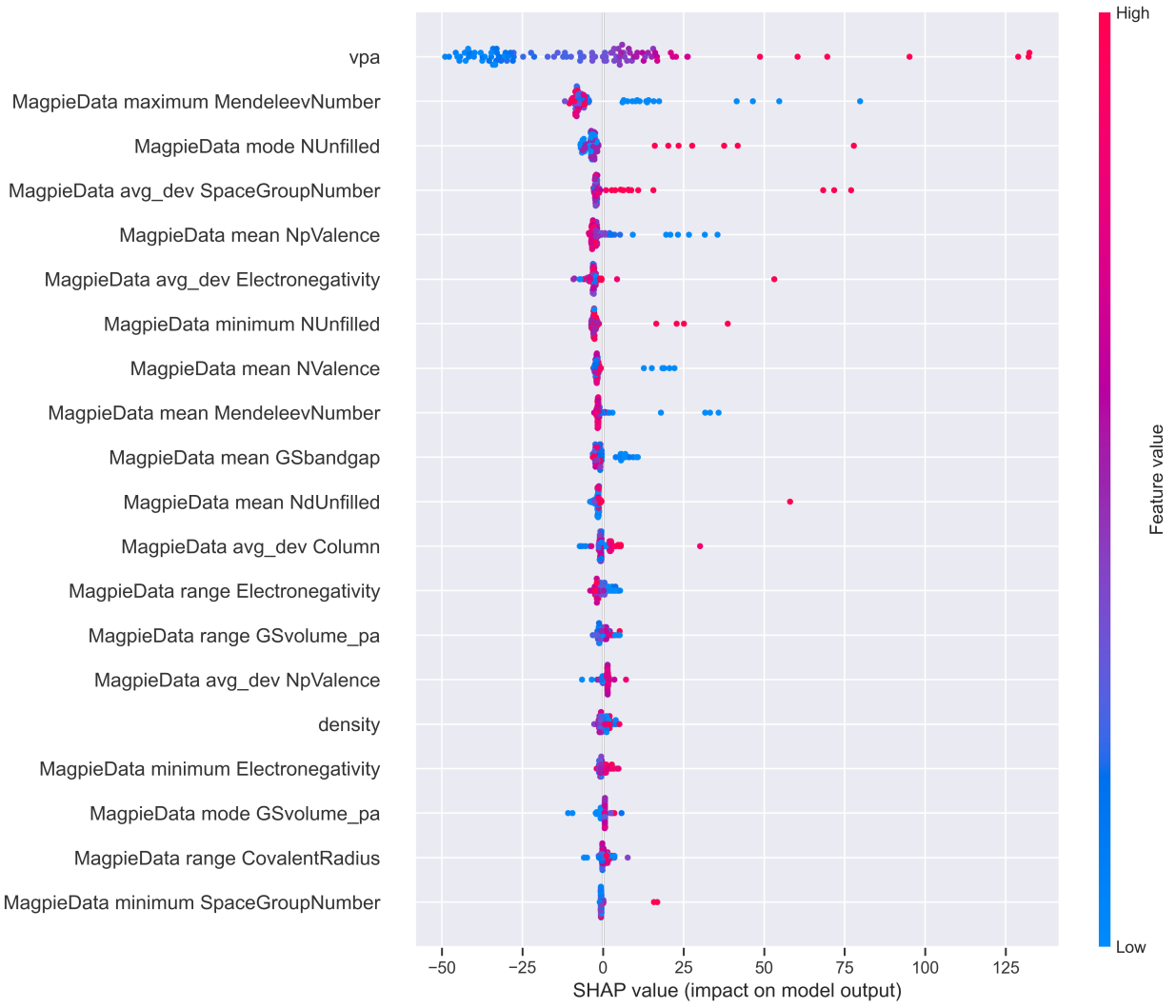

FIG. 11: SHAP summary plot. Each dot is a material; color indicates feature value (red = high, blue = low).

The concordance between SHAP importance and domain knowledge (vpa, bandgap, element identity) suggests the model exploits physically meaningful patterns rather than spurious correlations.

LIMITATIONS: HIGH-ENERGY PREDICTION FAILURE

While Random Forest performs well overall, limitations occur for materials with $E_{\text{exf}} >$ 300 meV/atom.

Error Analysis by Energy Range

TABLE V: Model MAE stratified by exfoliation energy range. All models exhibit catastrophic failure for $E_{\text{exf}} > 500$ meV/atom.

Energy Range (meV/atom)	**N Samples**	**Linear Reg. MAE (meV)**	**Random Forest MAE (meV)**	**Neural Net MAE (meV)**
0-100	99	48.2	**25.1**	30.0
100-200	19	60.7	**30.7**	33.1
200-300	3	53.2	**14.4**	75.6
300-400	**2**	191.2	**32.5**	130.3
400-500	**2**	131.1	**102.4**	130.4
500+	**2**	308.3	**329.6**	276.6

Root Cause: Severe Data Imbalance

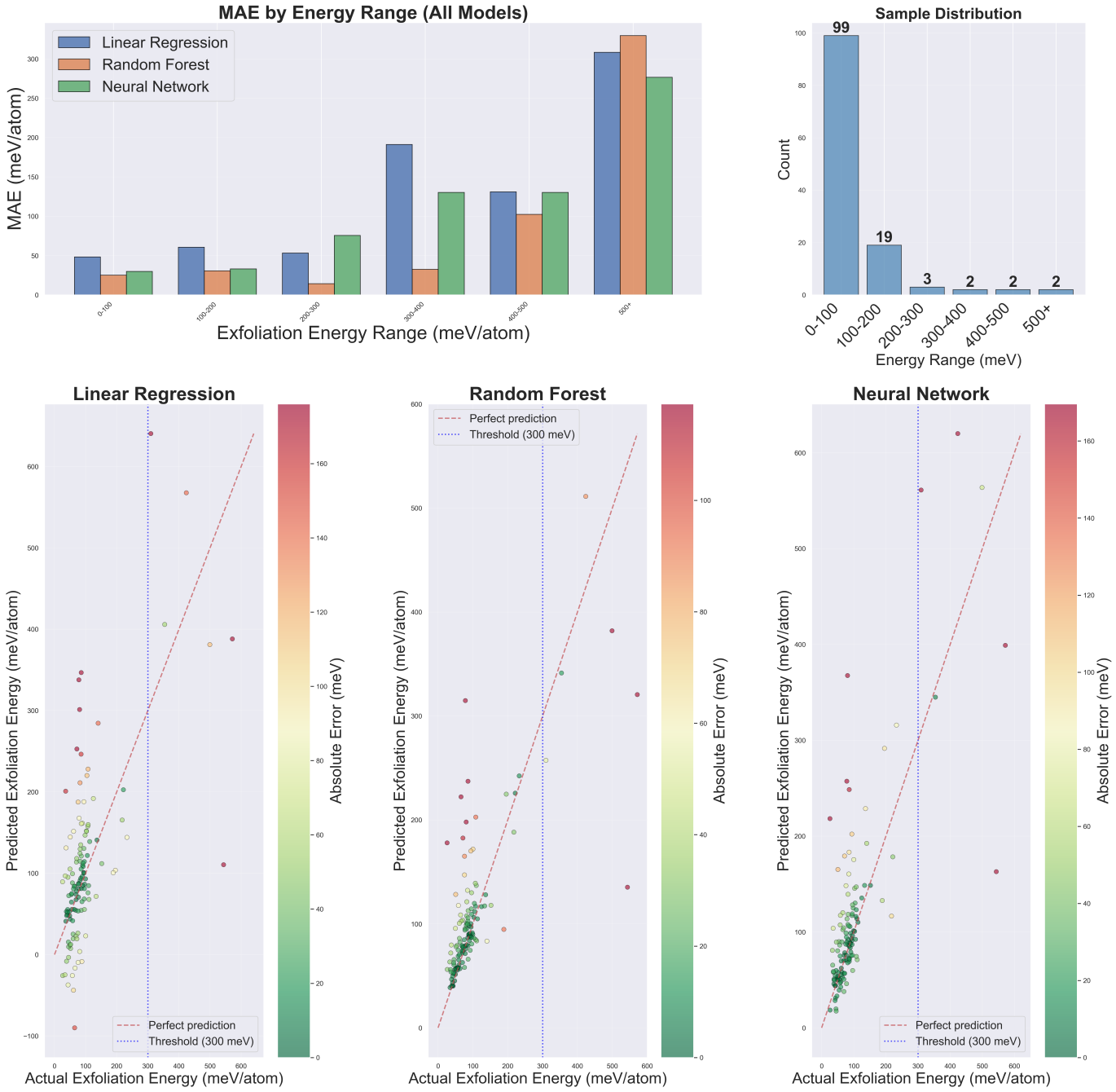

FIG. 12: High-energy prediction failure analysis. (Top left) MAE explodes for $E_{\text{exf}} > 300$ meV. (Top right) Sample distribution shows only 6/127 test materials exceed 300 meV—a 20:1 imbalance. (Bottom) Scatter plots colored by error magnitude highlight systematic underprediction above 300 meV threshold (blue dashed line).

Most materials in the dataset fall below 300 meV, with 121 samples (95%) in the low-energy regime and only 6 samples (5%) above this threshold, resulting in an extreme 20:1 imbalance.

This imbalance creates several issues. The MSE loss becomes biased toward the densely populated low-energy regime, effectively drowning out the sparse high-energy samples. High-energy materials also tend to occupy regions of feature space that are barely represented during training, meaning the model must extrapolate rather than interpolate. Most critically, with only a handful of high-energy examples available, the training set simply contains too little information for the model to learn the relevant structure–property relationships.

Comparison: Low vs. High Energy

TABLE VI: Performance degradation for high-energy materials ($E_{\text{exf}} > 300$ meV).

Model	**MAE ($E \leq 300$ meV)**	**MAE ($E > 300$ meV)**	**Degradation**
Linear Regression	50.3	210.2	4.2× worse
Random Forest	25.7	154.9	6.0× worse
Neural Network	31.6	179.1	5.7× worse

Attempted Mitigation: Sample Weighting

Reweighting high-energy samples 20× improved their MAE from 154.9 to 142.3 meV but degraded overall performance, confirming that weighting cannot overcome fundamental data scarcity with only 6 examples.

CONCLUSION

Summary of Findings

Random Forest achieved the strongest performance, reaching a test MAE of 31.84 meV/atom and improving upon the linear baseline by 45%. All models generalized well,

with train–test ratios between 0.96 and 1.08. The SHAP analysis confirmed that the models rely on physically meaningful descriptors such as volume per atom, bandgap features, and Mendeleev-number statistics. Despite these strengths, every model showed systematic underprediction above 300 meV/atom, reflecting the severe 20:1 imbalance in the dataset.

Limitations and Future Work

The primary limitations of this study stem from the scarcity of high-energy materials, the dependence on hand-crafted descriptors that may not capture subtle electronic effects, and the relatively small dataset size. Promising directions for future work include augmenting the dataset with additional high-energy materials from external databases [13], leveraging transfer learning from related materials-property models, using graph neural networks to learn physics-aware structural representations [14], and applying active learning to selectively expand the dataset where model uncertainty is highest.

Final Remarks

Machine learning for materials property prediction is not a replacement for physics-based simulation, but a *complementary accelerator*. By achieving a much faster convergence with acceptable accuracy for common materials, ML enables researchers to navigate the vast space of hypothetical compounds toward promising candidates worth deeper investigation. Future work should focus on hybrid workflows that leverage the strengths of both approaches: ML for breadth, DFT for depth.

I thank Professor Luciano Silvestri and the CMSE 492 teaching team for guidance throughout this project.

---

* filippo9@msu.edu

[1] K. S. Novoselov et al., "Electric Field Effect in Atomically Thin Carbon Films," *Science* **306**, 666 (2004).

[2] S. Z. Butler et al., "Progress, Challenges, and Opportunities in Two-Dimensional Materials Beyond Graphene," *ACS Nano* **7**, 2898 (2013).

[3] V. Nicolosi et al., "Liquid Exfoliation of Layered Materials," *Science* **340**, 1226419 (2013).

[4] A. Jain et al., "Commentary: The Materials Project: A materials genome approach to accelerating materials innovation," *APL Materials* **1**, 011002 (2013).

[5] J. Schmidt et al., "Recent advances and applications of machine learning in solid-state materials science," *npj Computational Materials* **5**, 83 (2019).

[6] A. Dunn et al., "Benchmarking materials property prediction methods: the Matbench test set and Automatminer reference algorithm," *npj Computational Materials* **6**, 138 (2020).

[7] G. Kresse and J. Furthmüller, "Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set," *Physical Review B* **54**, 11169 (1996).

[8] L. Ward et al., "Matminer: An open source toolkit for materials data mining," *Computational Materials Science* **152**, 60 (2018).

[9] M. Fernández-Delgado et al., "Do I Need Hundreds of Classifiers to Solve Real World Classification Problems?" *Journal of Machine Learning Research* **15**, 3133 (2014).

[10] G. Shwartz-Ziv and A. Armon, "Tabular data: Deep learning is not all you need," *Information Fusion* **81**, 84 (2022).

[11] L. Grinsztajn, E. Oyallon, and G. Varoquaux, "Why do tree-based models still outperform deep learning on typical tabular data?" *Advances in Neural Information Processing Systems* **35**, 507 (2022).

[12] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," *Advances in Neural Information Processing Systems* **30**, 4765 (2017).

[13] S. Haastrup et al., "The Computational 2D Materials Database: high-throughput modeling and discovery of atomically thin crystals," *2D Materials* **5**, 042002 (2018).

[14] T. Xie and J. C. Grossman, "Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties," *Physical Review Letters* **120**, 145301 (2018).

CODE AVAILABILITY

The complete code for this project is available at: https://github.com/AttackOnBreakfast/cmse492_project