Predicting Exfoliation Energy of 2D Materials Using Machine Learning

Artemiy Filippov*

Department of Computational Mathematics, Science and Engineering

Michigan State University, East Lansing, MI 48824

(Dated: November 3, 2025)

Abstract

This project predicts the exfoliation energy of two-dimensional (2D) materials using supervised machine learning. Exfoliation energy, measured in eV/atom, quantifies interlayer bonding strength and governs ease of isolating single layers. Density functional theory (DFT) gives accurate results but is computationally expensive. Using the MatBench JDFT-2D dataset and Matminer's Magpie/Density descriptors, I performed exploratory data analysis revealing nonlinear, heteroscedastic feature–target relations. Linear regression underfits, motivating nonlinear models. Future stages compare Random Forest and Neural Network regressors to evaluate nonlinear gains via MAE and RMSE. The goal is a reproducible, interpretable pipeline for fast materials screening.

BACKGROUND AND MOTIVATION

2D materials such as graphene and transition-metal dichalcogenides exhibit unique physical and electronic properties. The exfoliation energy—the energy to separate one atomic layer from the bulk—dictates synthesizability. While DFT provides accurate estimates, it is costly.

Machine learning offers fast approximations of DFT-level accuracy. Prior MatBench results show ML can capture structure–property relations via compositional and structural descriptors. This project builds ML regressors to predict exfoliation energy efficiently and interprets which physical features dominate.

DATA DESCRIPTION

The dataset is the **MatBench JDFT-2D** benchmark (Ward et al., 2020), derived from the Materials Project database. It contains ∼636 samples of layered compounds with exfoliation energies computed via DFT (vdW-optB88 functional).

Features: dozens of composition- and structure-based descriptors from Matminer's `MagpieData` and `DensityFeatures`. Target: exfoliation energy (eV/atom).

The energy distribution (Fig. 1) is strongly right-skewed—most materials are weakly bound (< 200 eV/atom), with few outliers above 1000 eV/atom. Correlation analysis (Fig. 2) shows weak linear trends; thus, nonlinear models are needed.

EXPLORATORY DATA ANALYSIS

Figures 1–4 summarize EDA and baseline performance. Scatter plots show weak, nonlinear dependencies on density and Mendeleev number. The baseline linear regression (Fig. 4) exhibits high error and no linear correlation, confirming model underfit.

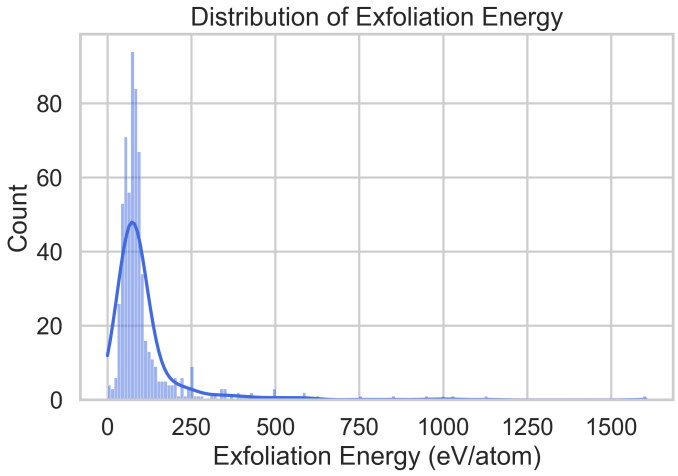

FIG. 1: Exfoliation-energy distribution: most samples below 200 eV/atom, with a long high-energy tail.

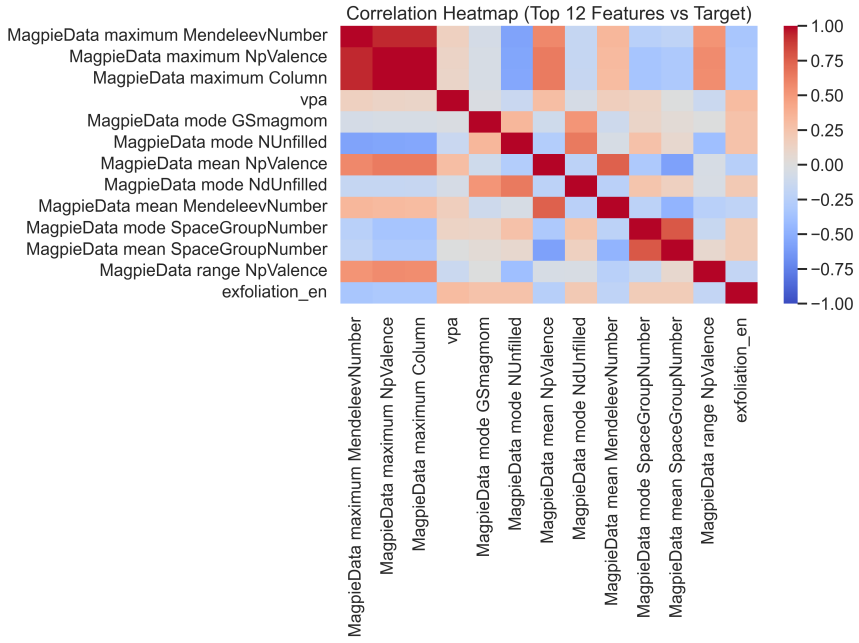

FIG. 2: Correlation heatmap for major compositional and structural descriptors. Nonlinear patterns dominate.

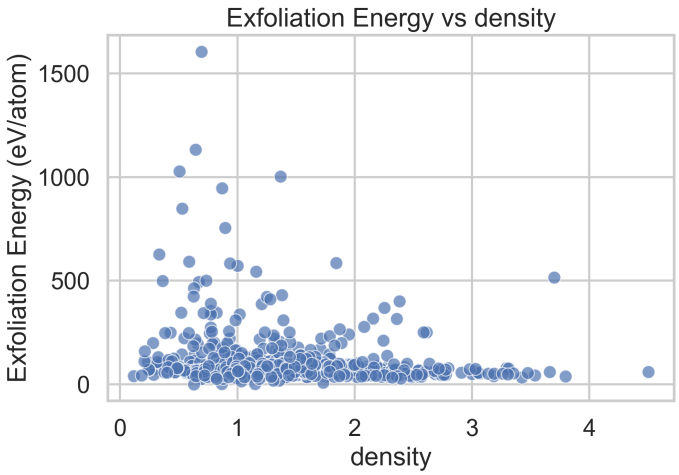

FIG. 3: Exfoliation energy vs. density showing nonlinear spread at low densities.

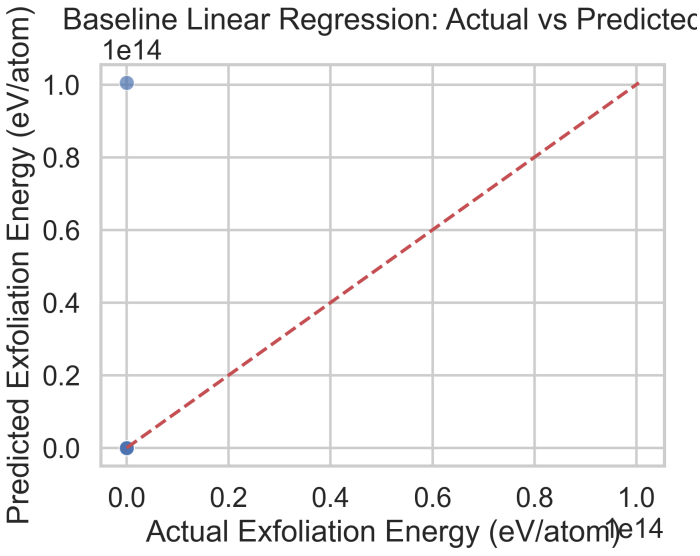

FIG. 4: Baseline linear regression: predicted vs. actual exfoliation energies. Poor correlation indicates underfit.

PROPOSED METHODOLOGY

The task is framed as supervised regression. Three models of increasing complexity will be benchmarked:

1. **Linear Regression** — interpretable baseline. 2. **Random Forest Regressor** — nonlinear ensemble with feature-importance insights. 3. **Feed-Forward Neural Network** — small MLP (2–3 layers, ReLU + Adam) for capturing nonlinear effects.

All models will be built with scikit-learn (and Keras for the NN) using 80/20 train–test splits and k-fold cross-validation.

EVALUATION FRAMEWORK

Performance metrics:

$$\text{MAE} = \frac{1}{n}\sum |y_i - \hat{y}_i|, \quad \text{RMSE} = \sqrt{\frac{1}{n}\sum (y_i - \hat{y}_i)^2}.$$

MAE reflects mean deviation; RMSE penalizes large errors. Success criterion: at least 50% MAE reduction relative to the baseline. Cross-validation and residual analysis will monitor generalization.

TIMELINE AND MILESTONES

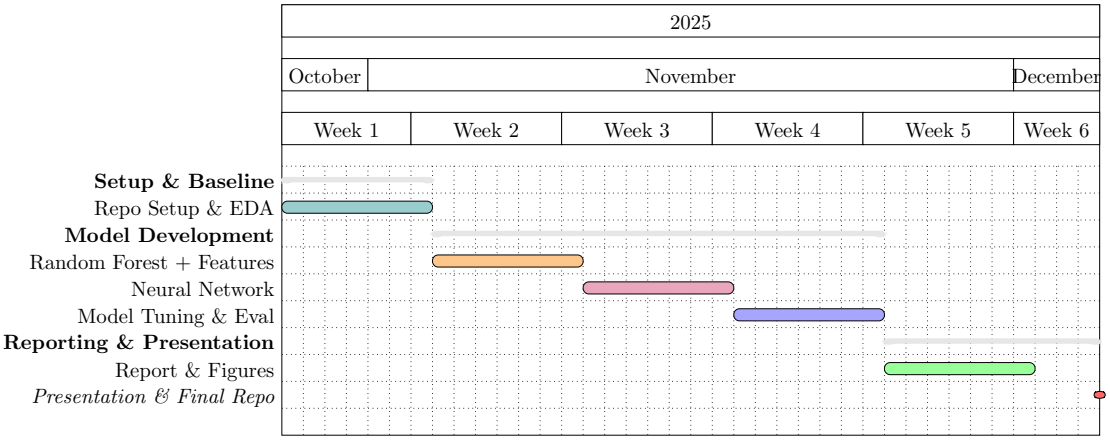

FIG. 5: Project timeline for CMSE 492, spanning late October–early December 2025. Thanksgiving week serves as buffer time for writing and figure refinement.

CONCLUSION

This project develops interpretable ML models for exfoliation-energy prediction. Combining linear, ensemble, and neural approaches enables a systematic study of model complexity vs. accuracy. Early baselines confirm nonlinear dependencies. The final stage will include SHAP-based feature importance, performance comparison, and code release on GitHub.

I thank Dr. Luciano Silvestri and the CMSE 492 teaching team for their guidance. All code and data are publicly available at: https://github.com/AttackOnBreakfast/cmse492_project.

---

* filippo37@msu.edu

[1] W. Ward et al., *npj Computational Materials*, 2020.

[2] A. Géron, *Hands-On Machine Learning*, O'Reilly Media, 2022.

[3] A. Jain et al., *APL Materials*, 2013.