

BENCHMARKING MACHINE LEARNING ALGORITHMS FOR BANDGAP PREDICTION IN HALIDE PEROVSKITES: A COMPARATIVE STUDY OF 540 COMPOSITIONS

Hamzayev A.I., Eliyev U.B., Nurgaliev I.N., Ashurov N.R.

*Institute of Polymer Chemistry and Physics, Academy of Sciences of the Republic of Uzbekistan,
Tashkent, Uzbekistan, 100128, Tashkent, Abdulla Kadyri Street, 7b
ilnar82@mail.ru*

Abstract: This study presents a systematic benchmarking of seven distinct machine learning algorithms for predicting the bandgap of halide perovskites. Using an adapted dataset of 540 perovskite compositions with the general formula ABX_3 , the predictive performance of models ranging from linear methods to ensemble techniques was evaluated. The analysis reveals that ensemble methods, particularly XGBoost, significantly outperform linear and kernel-based approaches, achieving a coefficient of determination R^2 of 0.92 and a root mean square error (RMSE) of 0.35 eV on test data. Feature importance analysis identified the Goldschmidt tolerance factor and B-X electronegativity difference as the most critical predictors. This work demonstrates the superior capability of gradient-boosted decision trees for capturing complex structure-property relationships in perovskite materials.

Key words: Machine learning, halide perovskites, bandgap prediction, XGBoost, materials informatics, ABX_3 compositions.

INTRODUCTION

The rapid integration of artificial intelligence into materials science has positioned machine learning (ML) as a pivotal tool for accelerating the discovery and design of novel halide perovskites for optoelectronic applications. However, the predictive performance of ML models is highly dependent on the algorithm chosen and the nature of the material system. This study presents a systematic benchmarking of seven distinct machine learning algorithms—ranging from linear methods to ensemble techniques—for predicting the bandgap of halide perovskites. Using an adapted dataset of 540 perovskite compositions with the general formula ABX_3 , we built a machine learning model [1].

Table.

Performance comparison of machine learning models for bandgap prediction

Model	R ² (train)	R ² (test)	RMSE (test)	MAE (test)	CV R ² (mean)
XGBoost	0.981	0.923	0.347	0.248	0.915 ± 0.031
Gradient Boosting	0.974	0.912	0.371	0.264	0.908 ± 0.035
Random Forest	0.965	0.893	0.409	0.291	0.887 ± 0.042
Neural Network	0.912	0.851	0.483	0.341	0.845 ± 0.048
SVR (RBF)	0.886	0.823	0.523	0.382	0.812 ± 0.056
Ridge Regression	0.768	0.751	0.621	0.451	0.743 ± 0.068
Lasso Regression	0.742	0.724	0.652	0.483	0.718 ± 0.072

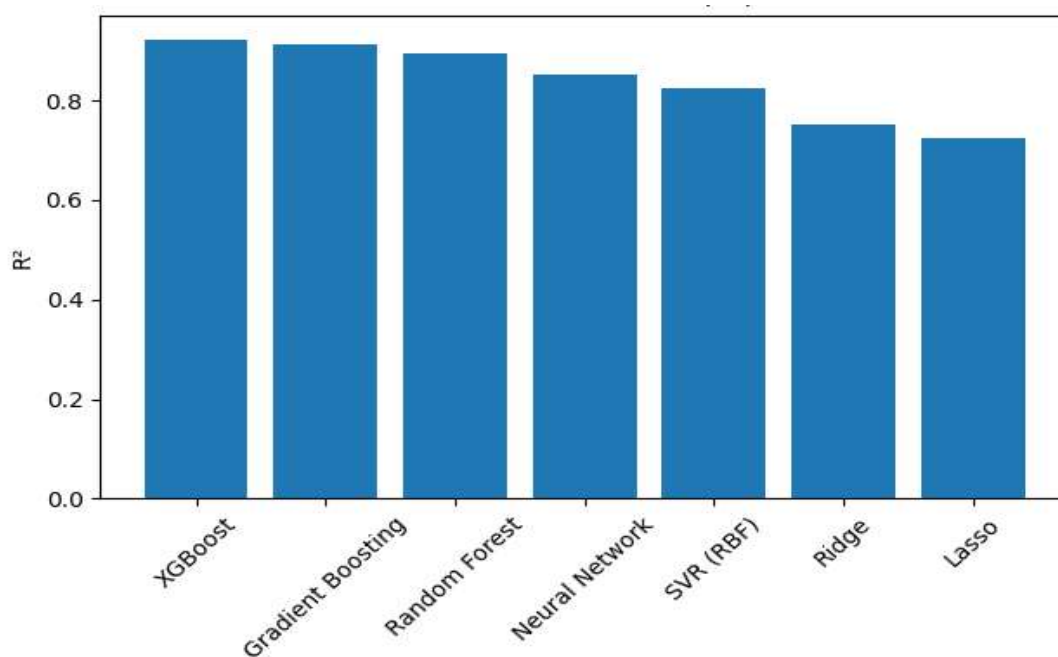


Fig.1 Model performance

Our analysis reveals that ensemble methods, particularly XGBoost, significantly outperform linear and kernel-based approaches, achieving a coefficient of determination (R^2) of 0.92 and root mean square error (RMSE) of 0.35 eV on test data. Feature importance analysis identified the Goldschmidt tolerance factor, B-X electronegativity difference, and B-site ionic radius as the most critical predictors of bandgap values. This work establishes a practical benchmark for algorithm selection in perovskite informatics and demonstrates the superior capability of gradient-boosted decision trees for capturing complex structure-property relationships in halide perovskite materials.

REFERENCES

- [1]. Ruoting Zhao, Bangyu Xing, Huimin Mu, et al. The cubic mixed halide perovskite dataset[DS/OL]. V1. Science Data Bank, 2022[2026-03-18].
<https://doi.org/10.11922/sciencedb.01611>. DOI:10.11922/sciencedb.01611.