ACCELERATED PHOTOCATALYTIC C–C COUPLING VIA INTERPRETABLE DEEP LEARNING: SINGLE-CRYSTAL PEROVSKITE CATALYST DESIGN USING FIRST-PRINCIPLES CALCULATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Photocatalytic C–C coupling reactions have garnered significant attention for their potential to drive sustainable chemical transformations. The design of efficient photocatalysts is critical in optimizing these reactions. In this study, we use a computational materials science approach, leveraging first-principles calculations to evaluate the bandgap values of 158 single-crystal perovskite materials. We employ a deep learning model, incorporating a multi-head-attention mechanism within a ResNet architecture, to predict the bandgap based on features such as τ, Group-A, Group-B, Pettifor number, χM-B, χP-B, Ea-A, cB, KB, and Ra-B. This model's performance is compared to traditional machine learning techniques, including K-means, MLP, Random Forest, PCA, and Multivariable Linear Regression. The results demonstrate that the self-attention ResNet model achieves a training R^2 of 0.819 and a test R^2 of 0.803, indicating strong predictive accuracy. The model's interpretability is enhanced by visualizing the permutation importance of each feature, shedding light on the contributions of various factors to the prediction. These findings highlight the potential of machine learning, particularly deep learning, in accelerating the design of photocatalysts for C–C coupling reactions.

1 INTRODUCTION

The photocatalytic C–C coupling reaction, which enables the efficient synthesis of carbon–carbon bonds, is a pivotal process in catalysis, especially for sustainable chemical production (Roy et al., 2023). The efficiency of photocatalysts directly impacts the rate and selectivity of this reaction. Perovskite materials have emerged as promising candidates for photocatalysis due to their tunable electronic properties, but designing effective catalysts requires a deep understanding of their structural and electronic characteristics. Traditional methods of catalyst design are often time-consuming and experimentally demanding (Wang et al., 2020). In this context, machine learning (ML) and deep learning (DL) techniques offer an accelerated approach to predicting material properties and optimizing catalyst performance (Ren et al., 2023; Li et al., 2024).

This paper explores the application of deep learning, specifically a multi-head-attention enhanced ResNet model, to predict the bandgap values of 158 single-crystal perovskite materials. The model's predictions are compared against traditional ML methods such as K-means, MLP, Random Forest, PCA, and multivariable linear regression. We focus on the interpretability of the model, making use of techniques such as feature importance analysis to provide insights into the decision-making process of the model (LeCun et al., 2015; He et al., 2016; Breiman, 2001; Hartigan & Wong, 1979; Voita et al., 2019; Ashish, 2017). This study focuses on a dataset of 158 single-crystal perovskite materials, carefully selected for their relevance to photocatalytic C–C coupling. The material features used to train our deep learning model were derived from an initial pool of 38 features through a rigorous feature engineering process, resulting in a set of 10 highly effective descriptors.

2 DATA COLLECTION

The dataset used in this study consists of first-principles computed bandgap values for 158 single-crystal perovskite materials. These materials were specifically chosen for their potential relevance to photocatalytic applications, focusing on a specific class of perovskites characterized by their similar crystal structures and elemental compositions conducive to specific electronic properties. While we initially explored larger databases such as the Materials Project, inconsistencies between some of their data and existing literature led us to construct a smaller, highly curated dataset using Density Functional Theory (DFT) calculations to ensure data accuracy. This approach allows for a more controlled and reliable dataset for training our machine learning models.

In addition to the bandgap values, we collected an initial set of 38 material features hypothesized to influence the electronic structure of the materials (Hafner, 2008; Wang et al., 2019; Hehre, 1976).. Through a process of feature engineering, we identified and selected the 10 most informative and efficient features as descriptors for our machine learning model. These features include (Green et al., 2014; Cheng et al., 2020; Wang et al., 2024):

- τ (transition metal parameter)
- Group-A (group of elements in the periodic table)
- Group-B (group of elements in the periodic table)
- Pettifor number (atomic bonding characteristics)
- χM-B (electronegativity of the metal-B component)
- χP-B (electronegativity of the perovskite-B component)
- Ea-A (activation energy for charge transfer)
- cB (bonding parameter)
- KB (bulk modulus)
- Ra-B (atomic radii)

The rationale behind selecting these 10 features is that they represent a combination of electronic, structural, and chemical properties known to significantly impact the bandgap of perovskite materials. Our feature engineering process involved analyzing feature correlations, assessing their individual and combined importance through preliminary model training, and considering domain knowledge from materials science. This selection aims to provide an efficient and physically meaningful representation of the perovskite materials for the predictive model.

3 DATA PREPROCESSING AND FEATURE ENGINEERING

Prior to model training, the dataset was preprocessed to handle missing values and normalize the features (Eck & Waltman, 2009). Features such as the transition metal parameters and atomic radii were scaled to ensure consistency across different material compositions. The dataset was then split into training and testing subsets, ensuring a proper distribution of data points. Feature engineering was performed to evaluate the relationships between different features and their impact on the bandgap values (Reitermanova et al., 2010).

4 MODEL DEVELOPMENT: MULTIHEAD-ATTENTION ENHANCED RESNET AND TRADITIONAL ML MODELS

The primary model used in this study is a deep learning model based on the ResNet architecture, enhanced with a self-attention mechanism to better capture the complex relationships between the features and the bandgap values (He et al., 2016; Voita et al., 2019; Ashish, 2017). The ResNet architecture was chosen due to its ability to learn deep representations of the data, while the self-attention mechanism improves the model's ability to focus on important features, enhancing its predictive power. The specific hyperparameters and training details of our ResNet model are as follows:

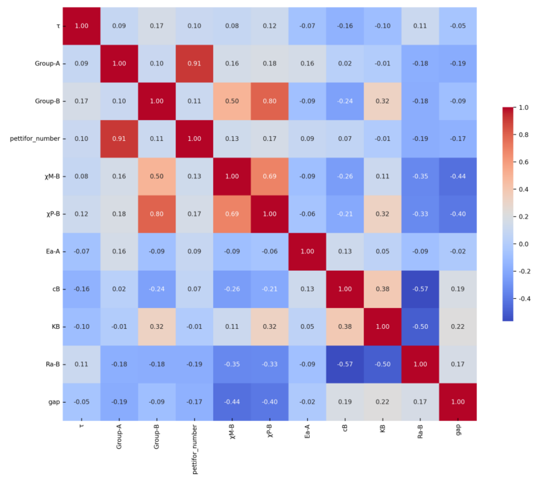

Figure 1: Pearson Heatmap showing the correlation between features.

- **Number of neurons in the hidden layers (neuro):** 1280
- **Number of ResNet layers (layer):** 10
- **Dropout rate (drop):** 0.4
- **Dimension of each attention head (head_dim):** 320
- **Number of training epochs (times):** 200
- **Learning rate (LR):** 0.0001
- **Learning rate decay factor (gam):** 0.999
- **Number of neurons in the transformer encoder layers (transneuro):** 160
- **Optimizer:** Adam

In addition to the ResNet model, several traditional machine learning models were also trained for comparison, including:

- K-means clustering (Hartigan & Wong, 1979)
- Multilayer Perceptron (MLP) (Taud & Mas, 2017)
- Random Forest (Breiman, 2001)
- Principal Component Analysis (PCA) (Abdi & Williams, 2010; Greenacre et al., 2022)
- Multivariable Linear Regression

5 Model Evaluation

The models were evaluated based on their performance on both the training and testing datasets. The primary evaluation metric used was the coefficient of determination (R^2), which measures the proportion of variance in the bandgap values explained by the model (Raschka, 2018; Eperon et al., 2016; Straus & Cava, 2022). The results of this evaluation are summarized in Table 1.

The ResNet model, with the added self-attention mechanism, achieved a training R^2 of 0.819 and a testing R^2 of 0.803, demonstrating its ability to generalize well to unseen data. Traditional models, such as Random Forest and MLP, performed adequately but did not match the performance of the deep learning model.

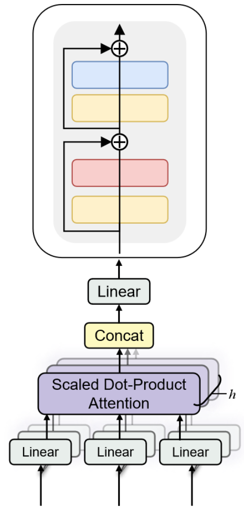

Figure 2: Architecture of the Multihead-Attention Enhanced ResNet model.

Table 1: Model performance comparison.

Model	Train R^2	Test R^2
ResNet	0.819	0.803
K-means	0.583	0.031
MLP	0.792	-66.437
Random Forest	0.946	0.581
PCA	0.439	-0.107
Linear Regression	0.005	-0.033

6 MODEL INTERPRETABILITY

To enhance the interpretability of the deep learning model, we performed a permutation importance analysis of the features. This analysis provided insight into how each feature contributes to the model's predictions. The permutation importance values were visualized in a boxplot (Figure 4), showing that certain features, such as the χM-B and Ea-A, had a significant impact on the model's performance (Altmann et al., 2010; Zhang & Zhu, 2018).

Additionally, partial dependence plots (PDPs) were generated for the Random Forest model, further shedding light on the relationships between features and the predicted bandgap. These plots demonstrate how variations in individual features influence the output of the model, with more pronounced curves indicating higher importance of the feature (Figure 5).

7 CONCLUSION AND OUTLOOK

This work demonstrates the potential of using deep learning, particularly a self-attention enhanced ResNet model, for the accelerated design of photocatalysts. The model's strong predictive performance and interpretability suggest that it can be used as a tool for discovering new materials with optimized properties for photocatalytic C–C coupling reactions. Future work will focus on expanding the dataset, exploring other deep learning architectures, and validating the predicted materials through experimental synthesis.

The development of this approach holds great promise for advancing the field of photocatalysis and for the broader application of AI in materials science.

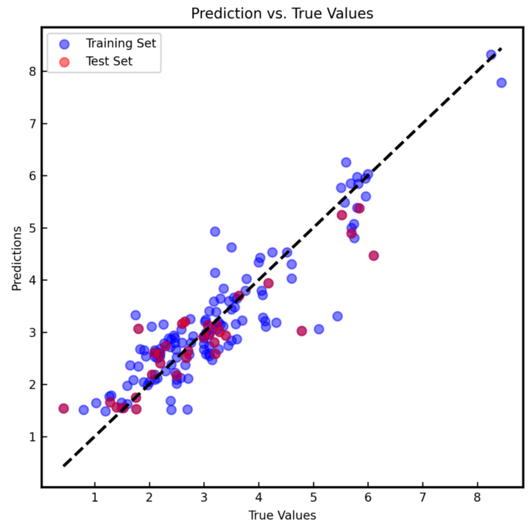

Figure 3: Predicted vs Observed bandgap values for the ResNet model.

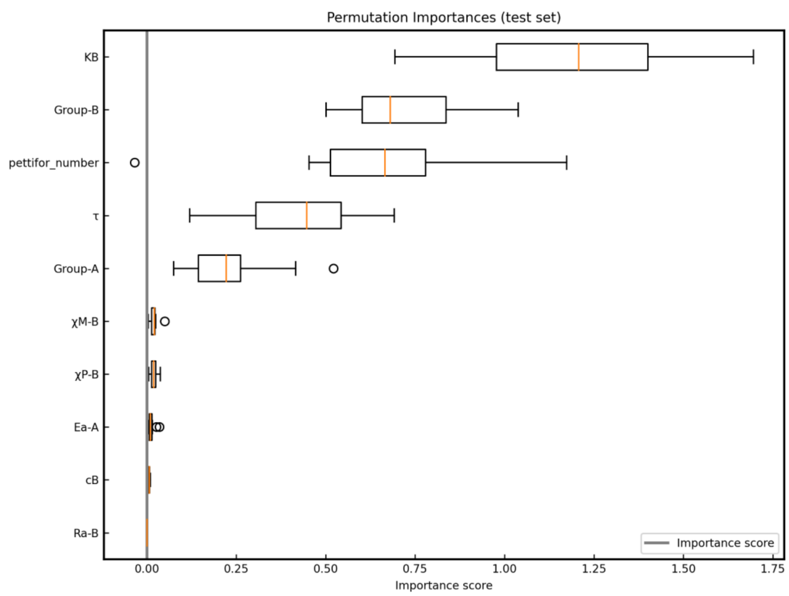

Figure 4: Permutation importance analysis of the features.

8 Data Availability

All code and data used in this study will be made publicly available on GitHub for reproducibility and further research purposes.

References

Hervé Abdi and Lynne J Williams. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459, 2010.

André Altmann, Laura Toloşi, Oliver Sander, and Thomas Lengauer. Permutation importance: a corrected feature importance measure. *Bioinformatics*, 26(10):1340–1347, 2010.

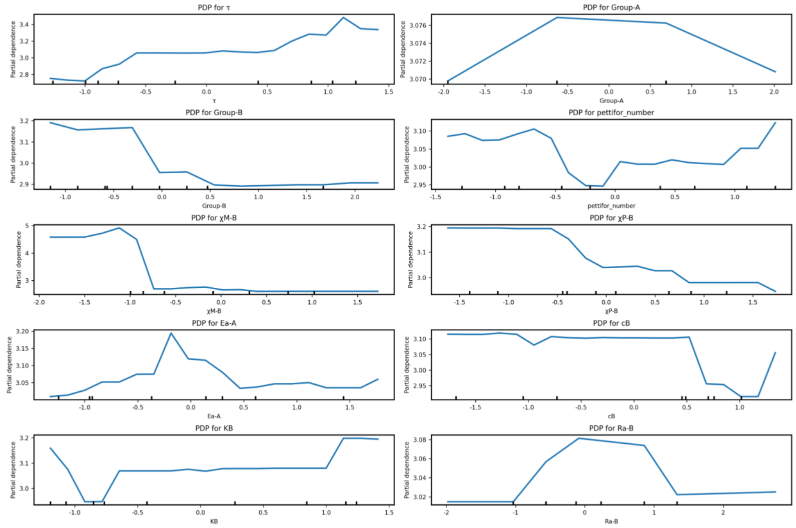

Figure 5: Partial Dependence Plots (PDPs) for the Random Forest model.

Vaswani Ashish. Attention is all you need. *Advances in neural information processing systems*, 30: I, 2017.

Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

Xiao Cheng, Shuang Yang, Bingqiang Cao, Xutang Tao, and Zhaolai Chen. Single crystal perovskite solar cells: development and perspectives. *Advanced Functional Materials*, 30(4):1905021, 2020.

Nees Jan van Eck and Ludo Waltman. How to normalize cooccurrence data? an analysis of some well-known similarity measures. *Journal of the American society for information science and technology*, 60(8):1635–1651, 2009.

Giles E Eperon, Tomas Leijtens, Kevin A Bush, Rohit Prasanna, Thomas Green, Jacob Tse-Wei Wang, David P McMeekin, George Volonakis, Rebecca L Milot, Richard May, et al. Perovskite-perovskite tandem photovoltaics with optimized band gaps. *Science*, 354(6314):861–865, 2016.

Martin A Green, Anita Ho-Baillie, and Henry J Snaith. The emergence of perovskite solar cells. *Nature photonics*, 8(7):506–514, 2014.

Michael Greenacre, Patrick JF Groenen, Trevor Hastie, Alfonso Iodice d'Enza, Angelos Markos, and Elena Tuzhilina. Principal component analysis. *Nature Reviews Methods Primers*, 2(1):100, 2022.

Jürgen Hafner. Ab-initio simulations of materials using vasp: Density-functional theory and beyond. *Journal of computational chemistry*, 29(13):2044–2078, 2008.

John A Hartigan and Manchek A Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, 28(1):100–108, 1979.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Warren J Hehre. Ab initio molecular orbital theory. *Accounts of Chemical Research*, 9(11):399–406, 1976.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

Haobo Li, Xinyu Li, Pengtang Wang, Zhen Zhang, Kenneth Davey, Javen Qinfeng Shi, and Shi-Zhang Qiao. Machine learning big data set analysis reveals c–c electro-coupling mechanism. *Journal of the American Chemical Society*, 146(32):22850–22858, 2024.

Sebastian Raschka. Model evaluation, model selection, and algorithm selection in machine learning. *arXiv preprint arXiv:1811.12808*, 2018.

Zuzana Reitermanova et al. Data splitting. In *WDS*, volume 10, pp. 31–36. Matfyzpress Prague, 2010.

Yuqing Ren, Yao Chen, Qingfei Zhao, Zhenmin Xu, Meijun Wu, and Zhenfeng Bian. Engineering palladium nanocrystals boosting c- c coupling by photocatalysis. *Applied Catalysis B: Environmental*, 324:122264, 2023.

Debojyoti Roy, Sunandita Paul, and Jyotishman Dasgupta. Photocatalytic terminal c- c coupling reaction inside water soluble nanocages. *Angewandte Chemie International Edition*, 62(45): e202312500, 2023.

Daniel B Straus and Robert J Cava. Tuning the band gap in the halide perovskite cspbbr3 through sr substitution. *ACS Applied Materials & Interfaces*, 14(30):34884–34890, 2022.

Hind Taud and Jean-Franccois Mas. Multilayer perceptron (mlp). In *Geomatic approaches for modeling land change scenarios*, pp. 451–455. Springer, 2017.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. *arXiv preprint arXiv:1905.09418*, 2019.

Kang Wang, Haipeng Lu, Xiaolin Zhu, Yixiong Lin, Matthew C. Beard, Yong Yan, and Xihan Chen. Ultrafast reaction mechanisms in perovskite based photocatalytic c–c coupling. *ACS Energy Letters*, 5(2):566–571, 2020. doi: 10.1021/acsenergylett.9b02714. URL `https://doi.org/10.1021/acsenergylett.9b02714`.

Shifu Wang, Fuhua Li, Jian Zhao, Yaqiong Zeng, Yifan Li, Zih-Yi Lin, Tsung-Ju Lee, Shuhui Liu, Xinyi Ren, Weijue Wang, et al. Manipulating cc coupling pathway in electrochemical co2 reduction for selective ethylene and ethanol production over single-atom alloy catalyst. *Nature Communications*, 15(1):10247, 2024.

Vei Wang, Nan Xu, Jin Cheng Liu, Gang Tang, and Wen-Tong Geng. Vaspkit: a pre-and post-processing program for vasp code. *arXiv preprint arXiv:1908.08269*, 2019.

Quan-shi Zhang and Song-Chun Zhu. Visual interpretability for deep learning: a survey. *Frontiers of Information Technology & Electronic Engineering*, 19(1):27–39, 2018.