Finding Perovskite Composites With Preferable Features: Simple ML algorithms

Gurgen Kolotyan[a], **Hayk Khachatryan**[a], **Arevik Asatryan**[a]

[a] *A.B. Nalbandyan Institute of Chemical Physics, Yerevan, Armenia 0014* goshkolotyan@gmail.com, hayk.khachatryan@gmail.com, arevik.asatryan@ichph.sci.am

1. Introduction

Perovskite materials are a promising class of semiconductors for next-generation solar cells and display materials due to their exceptional photovoltaic properties [1, 2]. These materials, typically based on organic - inorganic lead halides (e.g., $MAPbI_3$) and generally having the composition of ABX_3, exhibit high absorption coefficients, long carrier diffusion lengths, and tunable bandgaps, making them highly efficient light harvesters. Since their initial application in photovoltaics, perovskite solar cells (PSCs) have rapidly advanced, achieving high power conversion efficiencies > 25%, rivaling traditional silicon-based technologies. However, challenges such as crystal stability and finding the best performing composite remain key hurdles to commercialization, driving ongoing research into searching alternative compositions. With continuous innovation, perovskite solar cells hold great promise for revolutionizing renewable energy technologies.

2. Methods and Results

This paper has two primary focuses. First, it analyzes literature data to identify trends in material properties based on their elemental compositions. Second, it demonstrates how machine learning (ML) algorithms can effectively predict new composite features, reducing the need for time-consuming and resource-intensive chemical synthesis. The predicted results closely align with established literature trends, enabling us to select promising compositions tailored to specific applications and requirements. The study compares four ML algorithms to validate this predictive approach.

2.1 Related work

Machine learning (ML) approaches have been widely used to characterize composites and identify optimal structures [3, 4]. While specialized algorithms have been developed for composite mining [5], this paper demonstrates that perovskite mining can be effectively accomplished using both simple and sophisticated algorithms, with the choice depending on dataset size. Our results show that both approaches can yield reliable outcomes when properly matched to the data characteristics.

2.2 Results

Literature data shows that the ionic radius at each crystallographic site varies systematically with its chemical composition. At the A-site, the ionic radius reaches its maximum when Formamidinium (FA) content is highest, and its minimum when Cesium content is highest. Similarly, at the X-site, the ionic radius varies between its maximum with pure Iodide and its minimum with pure Chlorine content. These site-specific ionic radii directly affect the material's electronic properties, particularly the band gap energy (the energy difference between valence and conduction bands). As illustrated in Figure 2, increasing the X-site ionic radius leads to a non-linear decrease in band gap energy.

A key challenge in perovskite research is identifying compositions that combine both stability and high efficiency. To address this challenge, we applied data science and machine learning (ML) techniques to predict optimal material compositions. Our study analyzed multiple datasets extracted from research literature, comprising between 100 and 300 unique perovskite compositions, each with their corresponding structural and electronic parameters.

One of the major challenges in applying ML to perovskite prediction stems from data heterogeneity - both in dataset sizes and the variety of material parameters. These parameters include ionic radii, tolerance factors, and other structural and electronic descriptors. Depending on how these parameters are organized, ML algorithms identify different correlations, leading to varying predictive outcomes.

Our study evaluates how parameter distribution patterns affect prediction accuracy across different ML approaches. We validate selected predictions through density functional theory (DFT) calculations and experimental synthesis, demonstrating the practical value of data-driven

Finding Perovskite Composites With Preferable Features: Simple ML algorithms

Gurgen Kolotyan[a], **Hayk Khachatryan**[a], **Arevik Asatryan**[a]

[a] *A.B. Nalbandyan Institute of Chemical Physics, Yerevan, Armenia 0014* goshkolotyan@gmail.com, hayk.khachatryan@gmail.com, arevik.asatryan@ichph.sci.am

methods in accelerating the discovery of stable, efficient perovskite materials for solar cell applications.

We employed ML algorithms ranging from basic linear regression to sophisticated neural networks. To assess their effectiveness, we systematically compared their performance across different datasets, evaluating how well each algorithm adapted to the structure and organization of the available data. As shown in Figure 1, simpler ML approaches proved effective for smaller datasets, while Figure 2 demonstrates that more complex algorithms were better suited for larger datasets. This comparative analysis revealed key relationships between i) the ML model accuracy, ii) dataset size, iii) parameter distribution, iv) and feature relevance. By analyzing how prediction quality depends on the alignment between ML algorithms and data characteristics, we established guidelines for selecting optimal modeling approaches. These findings provide a framework for selecting suitable machine learning approaches based on the nature of the dataset, ultimately enhancing the efficiency of data-driven materials discovery.

2.3 Figures and tables

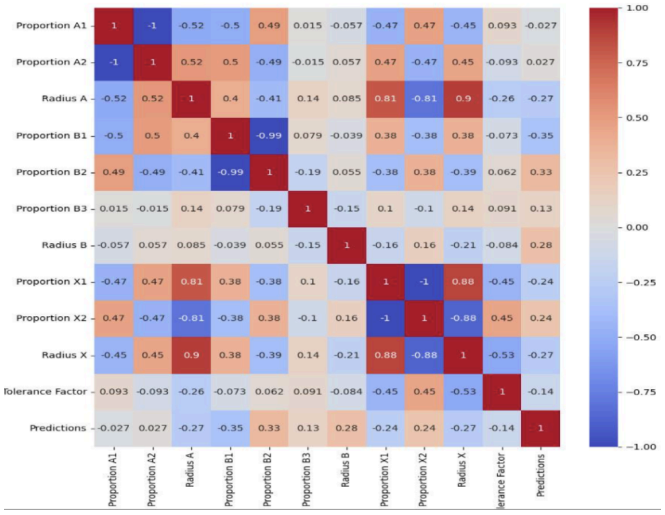

Fig. 1: Energy gap behaviour depending on anion radius - literature based.

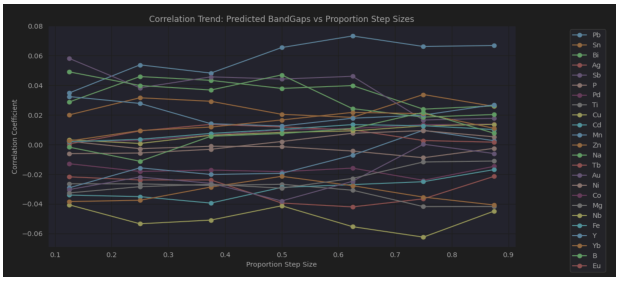

Fig. 2: Energy gap behaviour depending on anion radius and discrete cation radii - ML predicted.

2.4 Conclusions

In conclusion, our ML-based analysis provides a systematic method for identifying promising perovskite compositions. The selection process can be tailored by filtering compositions according to application-specific parameters. This data-driven approach provides a rational framework for identifying candidate materials for further experimental investigation, reducing the time and resources required for materials discovery.

Acknowledgments`

This work was supported by the Higher Education and the Science Committee of the Ministry of Education, Science, Culture and Sport RA [grant number 22RL-012].

References

[1] C. Zuo, H.J. Bolink, H. Han, J. Huang, D. Cahen, and L. Ding. *Advances in perovskite solar cells*. Pages 1500324. Advanced Science, 2016.

[2] M.A. Green, A. Ho-Baillie, and H.J. Snaith. *The emergence of perovskite solar cells.* Pages 506-514, Nature photonics, 2014.

[3] J. Kirman, A. Johnston, D. A. Kuntz, M. Askerka, Y. Gao, P. Todorović, D. Ma, G.G. Prive, and E. H. Sargent. *Machine-learning-accelerated perovskite crystallization.* Pages 938-947. *Matter,* 2020.

[4] Y. Buratti. *Machine learning for advanced characterisation of silicon solar cells.* PhD thesis., UNSW Sydney, 2023.

[5] A. Merchant, S. Batzner, S.S. Schoenholz., M. Aykol, G. Cheon, and E.D. Cubuk. *Scaling deep learning for materials discovery*. Pages 80-85. Nature, 2023.