HNUE JOURNAL OF SCIENCE
Natural Sciences 2026, Volume 71, Issue 1, pp. 54-64
This paper is available online at http://hnuejs.edu.vn/ns
DOI: 10.18173/2354-1059.2026-0005

MACHINE LEARNING ACCELERATED PROPERTY PREDICTION AND SCREENING OF STABLE LITHIUM-BASED BATTERY MATERIALS

To Anh Duc
Vietnam National Space Center, Vietnam Academy of Science and Technology, Hanoi, Vietnam
*Corresponding author: To Anh Duc, e-mail: taduc@vnsc.org.vn

Received February 17, 2026. Revised March 23, 2026. Accepted March 30, 2026.

Abstract. Rapid screening of battery materials is critical to meeting the growing demand for high-performance energy-storage solutions. In this study, we employ a data-driven approach to evaluate and characterize stable lithium-containing compounds with potential applications in next-generation batteries. We retrieved a dataset of 1,846 stable lithium-based materials from the Materials Project database using the mp-api. By leveraging the matminer library, we generated a comprehensive set of composition-based and structure-based features, including elemental- property and density descriptors. A Random Forest Regressor was trained to predict the band gap of these materials, a key electronic property that determines their suitability as electrolytes or electrodes. The model achieved a mean coefficient of determination (R^2) of 0.77 ± 0.04 and a mean absolute error (MAE) of 0.48 ± 0.05 eV across 10-fold cross-validation, demonstrating robust predictive capability. Feature-importance analysis revealed that electronegativity and Mendeleev number are the most significant predictors of band-gap energy. These findings underscore the effectiveness of machine learning in accelerating property prediction for battery materials and provide a pathway to more efficient experimental targeting.

Keywords: machine learning, battery materials, lithium, band-gap prediction, materials informatics.

1. Introduction

The global transition toward renewable energy and electric mobility has placed unprecedented demands on energy storage technologies [1]-[10]. Lithium-ion batteries (LIBs) have been the cornerstone of this revolution, powering applications ranging from portable electronics to grid-scale storage systems [11]-[20]. However, conventional LIBs, which rely on flammable organic liquid electrolytes, face inherent safety risks and energy density limitations. To overcome these challenges, the research community is

increasingly turning to solid-state batteries (SSBs), which utilize solid electrolytes to enable higher energy densities and improved safety profiles [21]-[30].

The successful deployment of SSBs depends on the discovery of novel materials with tailored electronic structures. Previous studies in materials informatics have demonstrated the power of machine learning (ML) for predicting the electronic band gap of inorganic solids [31]. Early work utilized support vector machines and random forests to screen broad chemical spaces, showing proving that composition-based descriptors can rival the accuracy of density functional theory (DFT) for band gap estimation [32]. More recently, advanced ensemble methods and gradient-boosted workflows have been applied to optimize feature selection for inorganic crystals, further refining our ability to identify insulators and semiconductors without the computational cost of ab initio simulations [33], [34].

Despite these advances, an important gap remains: while general band-gap models exist, there is a lack of high-throughput pipelines specifically tailored to the thermodynamic stability constraints of the lithium-containing chemical space required for solid-state electrolytes. Most large-scale screening studies do not explicitly filter for lithium-ion charge carriers in conjunction with absolute stability (energy above hull) [35]. This work addresses that gap by providing a targeted screening of 1,846 stable lithium compounds, specifically focusing on the electronic stability required for next-generation battery components.

By leveraging the Materials Project database and employing a Random Forest regression approach, we predict the electronic band gap using only computationally inexpensive descriptors. We detail our data acquisition strategy, feature-engineering workflow, and rigorous model evaluation, and provide insight into the key chemical drivers of electronic structure in battery materials.

2. Content

2.1. Methodology

2.1.1. Data acquisition

Data was retrieved from the Materials Project (MP) database using the mp-api client (v0.41.1). The search query was designed to filter for materials that satisfy three criteria: (1) containing the element Lithium ("Li"), (2) being thermodynamically stable (energy above hull = 0 eV/atom), and (3) having a calculated band gap. The specific fields requested for each entry were material_id, formula_pretty, structure, band_gap, and energy_above_hull. This yielded a final dataset of 1,846 unique crystal structures, encompassing a diverse range of chemistries including oxides, sulfides, phosphides, and halides.

2.1.2. Feature engineering

To convert the raw crystal structures into rigorous numerical descriptors suitable for machine learning, we used the matminer library. We generated a hybrid feature set that combines compositional and structural information:

Compositional features (Magpie): We employed the ElementProperty.from_preset ("magpie") featurizer. This algorithm computes a suite of statistics for atomic properties (atomic number, atomic mass, melting point, boiling point, electronegativity, etc.), stoichiometry-weighted statistics. For each property, the following statistics were calculated: minimum, maximum, range, mean, average deviation, and mode. This produced a comprehensive descriptor vector capturing the chemical nature of the constituent elements.

Structural features (Density): The DensityFeatures featurizer was used to extract geometric information, specifically density (g/cm³), volume per atom (A^3/atom), and packing fraction. These features serve as proxies for crystal-lattice openness, which is often correlated with electronic properties and ion mobility.

The resulting feature matrix was cleaned by removing any entries with missing values (NaN) generated during the featurization process (e.g., due to missing elemental data for rare earth elements).

2.1.3. Model training and evaluation

We employed the Random Forest Regressor, an ensemble learning method constructed from a multitude of decision trees, implemented via scikit-learn. To ensure robustness, the model was optimized via grid search ($n\_estimators = 500, max\_depth = 20$) and its performance was evaluated using 10-fold cross-validation on the full dataset of 1,846 materials using three standard metrics: coefficient of determination (R^2), mean absolute error (MAE), and root mean squared error (RMSE).

2.2. Results and discussion

2.2.1. Model performance

The Random Forest model demonstrated strong predictive capabilities on the held-out test set, primarily driven by the rich information contained within the compositional and structural features.

Figure 1 shows the parity plot comparing the model's predicted band gaps against the ground-truth DFT values. A tight clustering of data points around the red 1:1 ideal line is observed, indicating high model accuracy. Across the 10-fold cross-validation, the optimized model achieved a mean coefficient of determination (R^2) of 0.77 ± 0.04, indicating that the model successfully captures 77% of the variance in the band gap data. The mean absolute error (MAE) of 0.48 ± 0.05 eV provides a practical and tight bound on the prediction uncertainty. The low standard deviation across the 10 folds confirms that the model is stable and effectively generalizes across the diverse chemical space without overfitting. This is a significant result for a model relying on computationally inexpensive descriptors, as it suggests that the majority of the electronic structure information is encoded within simple composition and density metrics. For a screening workflow, this error is acceptable: missing a precise band gap by approximately 0.5 eV is unlikely to misclassify a metal as a wide-bandgap insulator, thereby facilitating the effective prioritization of potential solid electrolyte candidates for further investigation.

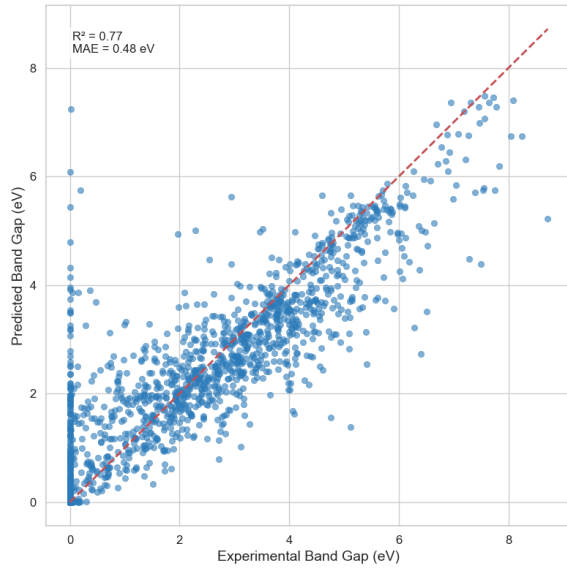

Figure 1. Parity plot of predicted vs. actual band gaps

Given that solid electrolytes typically require a band gap exceeding 3 eV to ensure electrochemical stability, an error of ~0.5 eV allows the model to effectively filter out the significant population of metallic and narrow-gap materials in the dataset, thereby prioritizing only high-probability candidates for subsequent, more computationally intensive DFT verification. To examine the reliability of the model across the property range, we analyzed the residuals (Figure 2). The residuals, defined as $(y_{actual} - y_{predicted})$, are plotted against the predicted band gap. Ideally, these should be randomly distributed around zero with no discernible pattern. The plot confirms a largely homoscedastic behavior, meaning the error variance is relatively constant. There is no significant systematic bias where the model consistently over- or under-predicts for small or large band gaps. This randomness in the error distribution validates the assumption that the Random Forest model has learned the underlying generalizable trends rather than overfitting to specific material classes.

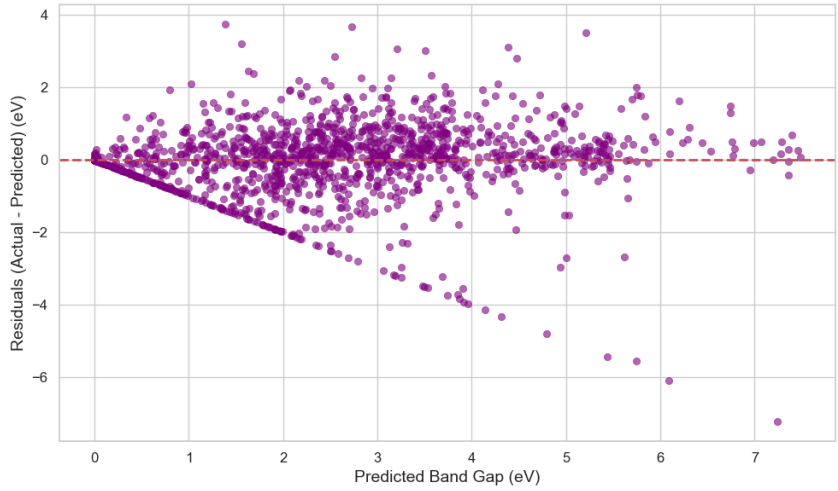

Figure 2. Residual plot showing the distribution of prediction errors

2.2.2. Feature-importance analysis

Understanding why a model makes a prediction is as important as the prediction itself, especially in materials science, where physical intuition is paramount.

Figure 3 displays the top 10 most important features identified by the Random Forest model, ranked by their mean decrease in impurity (MDI). The results are striking: features related to Electronegativity (e.g., maximum difference, range, mean) and Mendeleev Number dominate the list. This aligns perfectly with fundamental chemical theory. The band gap of a material is fundamentally the energy difference between the valence band maximum (bonding states) and conduction band minimum (anti-bonding states). This energy splitting is governed by the ionicity of the bonds: highly ionic bonds (large electronegativity difference between cation and anion) lead to strong localization of electrons and consequently wider band gaps (e.g., LiF). Conversely, covalent or metallic bonding (small electronegativity difference) results in delocalized electrons and smaller or zero band gaps. The model's feature importance ranking aligns with established chemical principles regarding the role of ionicity in electronic structure.

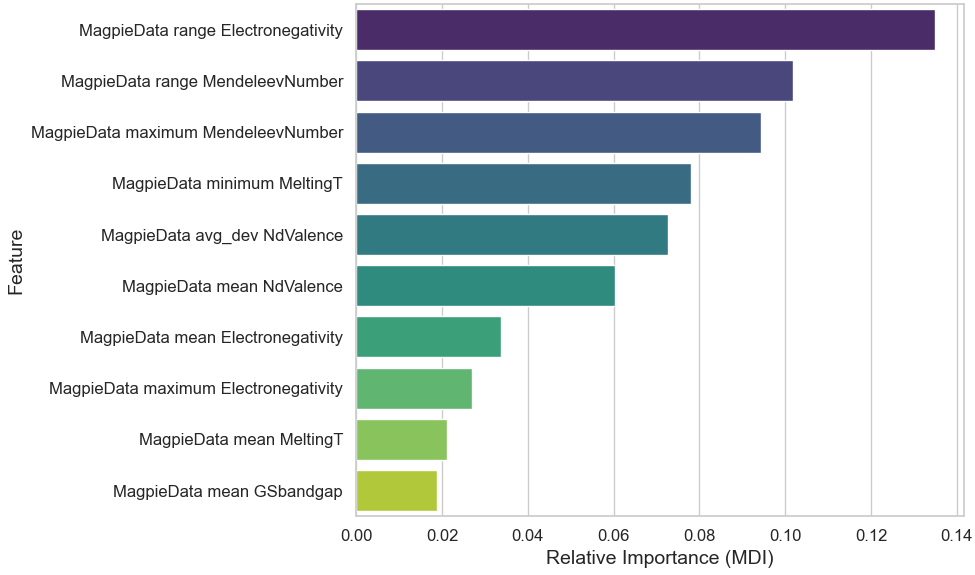

Figure 3. Top 10 features ranked by mean decrease in impurity (MDI)

The significant predictive power of the Mendeleev number, appearing in both its maximum and range statistics, stems from its ability to capture chemical similarity more effectively than the standard atomic number. While electronegativity-based features dominate the model by quantifying the energy splitting between bonding and anti-bonding states, the Mendeleev number provides a heuristic ordering that groups elements with similar chemical 'personalities' and structural preferences. This allows the Random Forest model to distinguish among different crystalline environments and coordination geometries that, while having similar ionicity, result in distinct electronic band gaps. To explicitly visualize this discovered relationship, Figure 4 plots the actual band gap against the single most important feature: the range of electronegativity in the compound. A clear, positive correlation is visible. As the difference in electronegativity between the constituent elements increases, the band gap tends to increase. This trend explains why

Li-halides and oxides (high electronegativity difference) are insulators, while Li-rich intermetallics are conductors. The color gradient, representing the band gap magnitude, further emphasizes this transition from metals (bottom left) to insulators (top right).

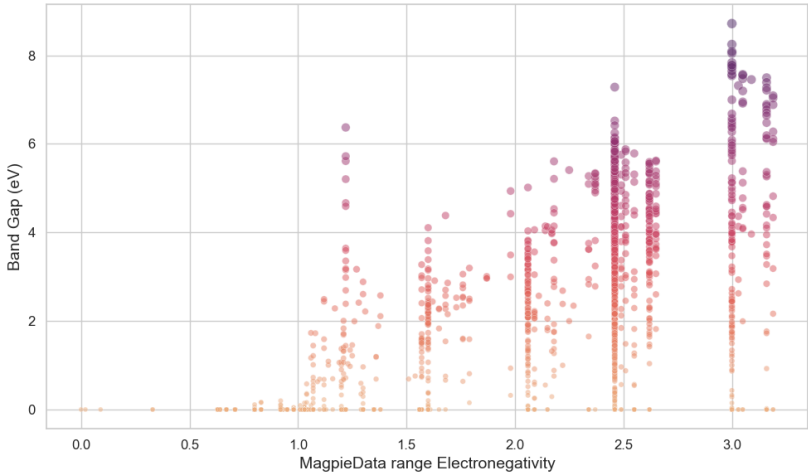

Figure 4. Scatter plot of band gap vs. electronegativity range

2.2.3. Dataset characteristics

Finally, we analyze the composition and distribution of the dataset itself to understand the domain of applicability of our model.

Figure 5 illustrates the distribution of band gaps within our stable lithium material dataset. The distribution is bimodal, characterized by a large peak near 0 eV (metals/conductors) and a broad tail extending to >8 eV (insulators). This distribution is typical for inorganic databases. The significant population of materials with band gaps > 3 eV is encouraging for solid-state battery research, as these are the prime candidates for solid electrolytes. The model's ability to distinguish these high-gap materials from the metallic clutter is its primary value proposition.

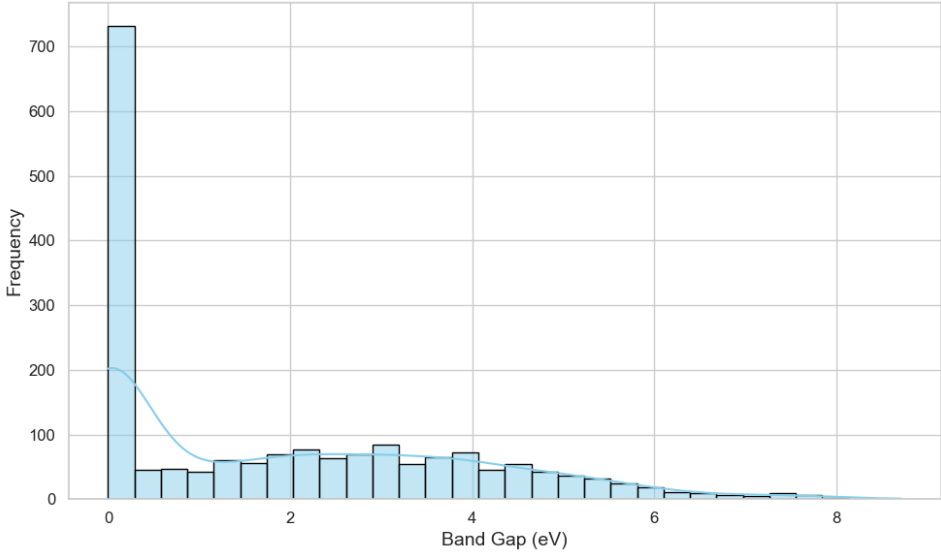

Figure 5. Histogram of band gap values

Chemical Diversity Figure 6 breaks down the elemental composition of the dataset, showing the 15 most frequent elements (excluding Li). Quantitatively, the dataset is dominated by oxides (approximately 38%), followed by phosphates (10%), fluorides (9%), and sulfides (6%), ensuring that the model is trained on the most technologically relevant chemical spaces for modern battery components. Oxygen is by far the most prevalent anion, followed by Sulfur, Phosphorus, and Fluorine. This reflects the intense research focus on oxides (classic cathodes like $LiCoO_2$), sulfides (superionic conductors like $Li_{10}GeP_2S_{12}$), and phosphates (olivine LFP). The presence of these elements confirms that our dataset covers the chemically relevant space for modern battery technology. Any model trained on this data is thus directly applicable to the most promising material families currently under investigation.

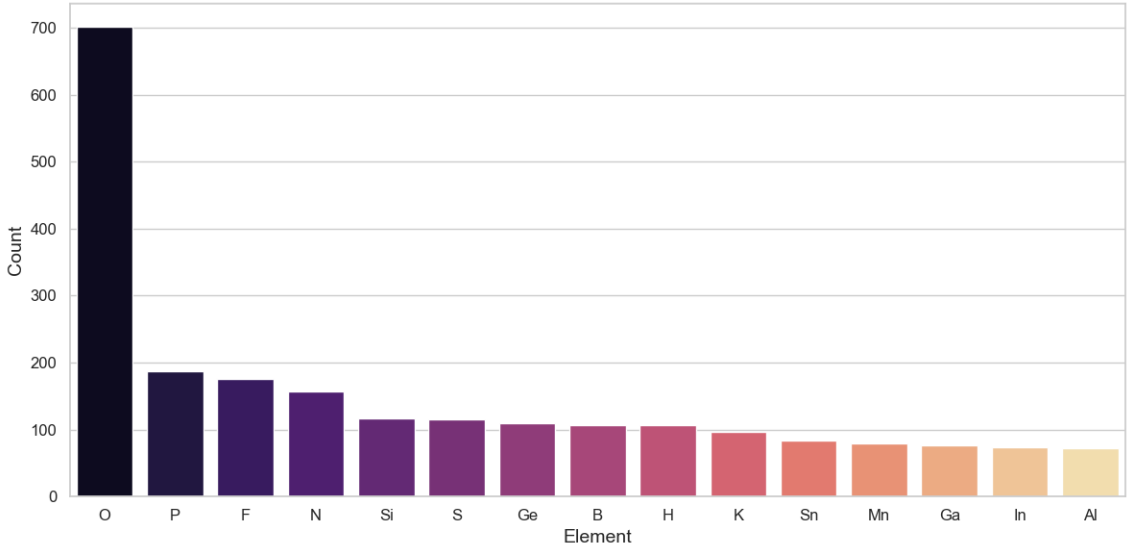

Figure 6. Frequency of elements in the dataset

To elucidate the interrelated dependencies within the feature space, a Pearson correlation heatmap (Figure 7) was generated for the 10 most important descriptors and the experimental band gap. The strongest linear relationship with the band gap is observed in the electronegativity range ($r = 0.68$), confirming a fundamental link between elemental polarity contrast and resultant electronic structure. Furthermore, significant multicollinearity is present among several secondary features, particularly between different variants of the Mendeleev number and electronegativity. This strong inter-feature dependence underscores the rationale for employing robust, nonlinear ensemble models like the Random Forest Regressor, which can naturally disentangle and appropriately weight complex, interrelated chemical spaces without being limited by strict linear assumptions.

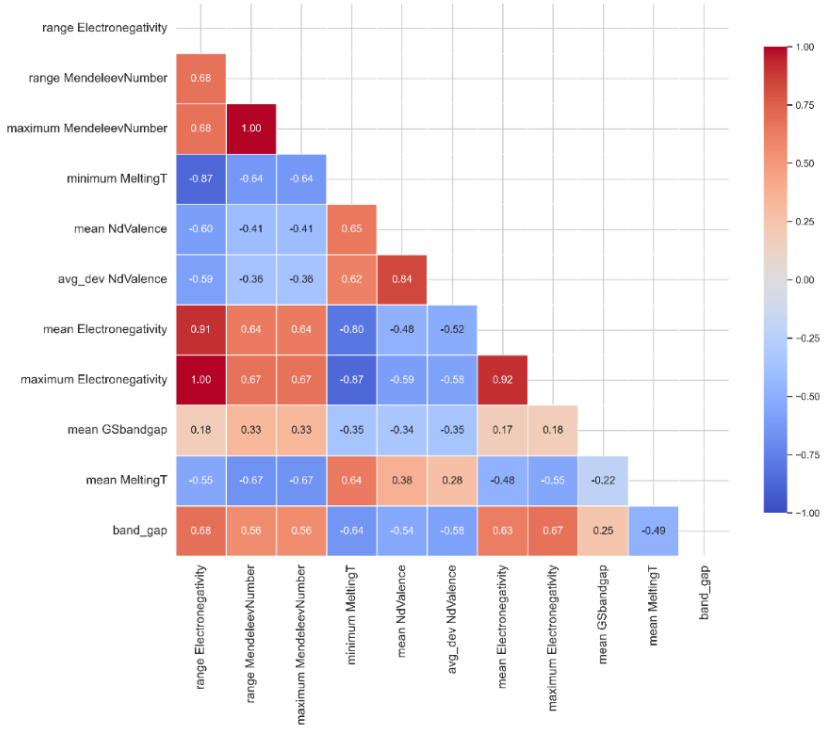

Figure 7. Lower-triangle correlation heatmap of the top 10 features and band gap

To further contextualize the physical and chemical generalization captured by the model, Figure 8 visualizes the underlying dataset distributions for the most influential descriptors identified through mean decrease in impurity: Electronegativity range, Mendeleev Number range, and maximum Mendeleev Number. The violin plots for these top predictors clearly demonstrate profound, continuous property variance across the 1,846 lithium-based candidate materials. Specifically, exhibiting such broad, overlapping distributions across Mendeleev-based features, which encode atomic scaling rules and structural clustering alongside electronegativity, confirms the rich chemical diversity of the dataset. This widespread variability provides a stabilizing basis for cross-validation, confirming that the Random Forest model captures generalizable physicochemical principles from highly heterogeneous compositional spaces rather than rigidly overfitting to narrow, overrepresented subclasses of materials.

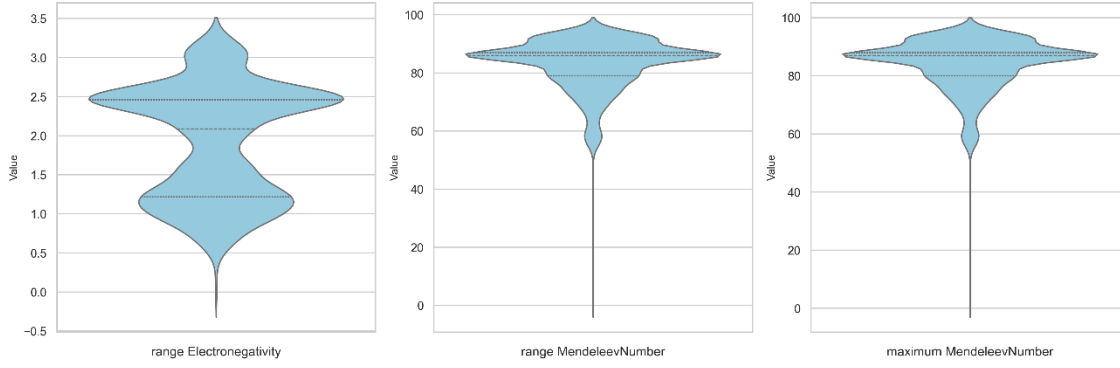

Figure 8. Feature distributions of the top 3 most important features

3. Conclusions

We have successfully developed and validated a machine-learning framework for the rapid prediction of electronic band gaps in stable lithium-based battery materials. By leveraging a high-quality dataset from the Materials Project and employing rigorous feature-engineering strategies, our Random Forest model achieved strong predictive accuracy, with a mean $R^2 = 0.77 \pm 0.04$ and a mean MAE of 0.48 ± 0.05 eV, as validated by 10-fold cross-validation. This level of accuracy is sufficient to distinguish among metals, narrow-bandgap semiconductors, and wide-bandgap insulators, making it a valuable tool for high-throughput screening.

The feature-importance analysis provided critical physical insights, identifying electronegativity and Mendeleev Number as the primary determinants of band-gap magnitude. This reinforces the chemical intuition that the ionicity of bonds and the periodic character of constituent elements dictate the electronic structure. Importantly, our model relies only on computationally inexpensive descriptors, composition, and density, thereby bypassing the need for costly DFT calculations during the initial screening phase. This capability represents a significant acceleration in the materials-discovery pipeline.

Furthermore, residual analysis confirms the robustness of the model, showing no systematic bias across the range of band gaps. The identification of a diverse set of stable lithium compounds, including oxides, phosphates, and sulfides, highlights the richness of the available chemical space for optimization. As the demand for safer and higher-energy density batteries intensifies, data-driven approaches like the one presented here will become increasingly indispensable. They not only accelerate the identification of promising candidates for solid-state electrolytes and cathodes but also provide a data-driven framework that supports and extends existing physical intuition.

To further refine this predictive framework, future efforts will integrate more sophisticated structural representations such as Voronoi tessellation and graph-based embeddings (e.g., graph neural networks) to better capture local chemical environments. Additionally, we aim to extend the pipeline to perform multi-objective optimization, simultaneously predicting ionic conductivity, oxidation stability, and shear modulus to design truly optimal solid-state battery materials.

REFERENCES

[1] Panesar A & Sampson O, 2026. Environmental impact of energy storage technologies and future renewable grids. *Journal of Energy Storage*, 152, 120501.

[2] Khan MFH, Biswas AK, Ahmed I, Shovon SM, Akash FA, Rahman A & et al., 2026. Techno-economic analysis and life cycle assessment of energy storage technologies: A comprehensive review. *International Journal of Green Energy*, 23(3), 535-559.

[3] Vallese L, Javadi H, Badenes B, Urchueguia JF, Lombardo G, Menegazzo D & et al., 2026. A comprehensive review of thermal energy storage technologies and their applications: Creation of a database. *Renewable and Sustainable Energy Reviews*, 225, 116133.

[4] Guo J, Jing Y, Hou W, Wang T, Ma S & He G, 2024. Demands and challenges of energy storage technology for future power systems. *Energy Internet*, 1(2), 116-122.

[5] Jiang T, Shen D, Zhang Z, Liu H, Zhao G, Wang Y, et al., 2025. Battery technologies for grid-scale energy storage. *Nature Reviews Clean Technology*, 1(7), 474-492.

[6] Helwig A & Bell J, 2024. What energy storage technologies will Australia need as renewable energy penetration rises?. *Journal of Energy Storage*, 95, 112701.

[7] Xu HJ, Han XC, Hua WS, Friedrich D, Santori G, Bevan E & et al., 2025. Progress on thermal storage technologies with high heat density in renewables and low carbon applications: Latent and thermochemical energy storage. *Renewable and Sustainable Energy Reviews*, 215, 115587.

[8] Li Z & Deusen D, 2025. Role of energy storage technologies in enhancing grid stability and reducing fossil fuel dependency. *International Journal of Hydrogen Energy*, 102, 1055-1074.

[9] Waseem M, Lakshmi GS, Ahmad M & Suhaib M, 2025. Energy storage technology and its impact in electric vehicle: Current progress and future outlook. *Next Energy*, 6, 100202.

[10] Aghmadi A & Mohammed OA, 2024. Energy storage systems: Technologies and high-power applications. *Batteries*, 10(4), 141.

[11] Baskaran D, Rajeswari S, Kanimozhi K & Byun HS, 2026. Bibliometric insights into lithium-ion battery research and recycling trends (2011–2022). *International Journal of Energy Research*, 2026(1), 4339273.

[12] Degen F, Winter M, Bendig D & Tübke J, 2023. Energy consumption of current and future production of lithium-ion and post lithium-ion battery cells. *Nature Energy*, 8(11), 1284-1295.

[13] Marri GK, Ee YJ, He Z & Ho JY, 2026. Recent advancements in internal and external thermoregulation strategies for lithium-ion batteries. *Renewable and Sustainable Energy Reviews*, 225, 116127.

[14] Korde VB, Khelkar AB, Khot S, Malavadakar P, Deshmukh P & Amalraj S, 2026. Advancements of lithium-ion battery recycling: Transitioning from traditional methods to AI and machine learning techniques. *Renewable and Sustainable Energy Reviews*, 225, 116180.

[15] Yang Q, Su Y, Wang X, Xiao K, Chen T & Fu J, 2026. Emerging strategies for sustainable holistic recycling of spent lithium-ion batteries. *Advanced Functional Materials*, 36(4), e14524.

[16] Fu S, Fan H, Jin Z, Ji F, Tao Y, Dong Y & et al., 2026. Recent progress in state of health estimation for lithium-ion batteries: From laboratory to practical application. *Renewable and Sustainable Energy Reviews*, 226, 116323.

[17] Zhang S, Liu Y, Chen G, El-Bahy ZM, Alshammari DA, Helal MH & et al., 2026. Adaptive localized ectopic structure enhances regeneration of spent lithium-ion battery cathodes. *Advanced Materials*, 38(1), e13547.

[18] Madani SS, Shabeer Y, Allard F, Fowler M, Ziebert C, Wang Z, et al., 2025. A comprehensive review on lithium-ion battery lifetime prediction and aging mechanism analysis. *Batteries*, 11(4), 127.

[19] Yao YX, Xu L, Yan C & Zhang Q, 2025. Principles and trends in extreme fast charging lithium-ion batteries. *EES Batteries*, 1(1), 9–22.

[20] Gong T, Duan X, Shan Y & Huang L, 2025. Gas generation in lithium-ion batteries: Mechanisms, failure pathways, and thermal safety implications. *Batteries*, 11(4), 152.

[21] Alkhalidi A, Khawaja MK & Ismail SM, 2024. Solid-state batteries, their future in the energy storage and electric vehicles market. *Science Talks*, 11, 100382.

[22] Thomas F, Mahdi L, Lemaire J & Santos DM, 2024. Technological advances and market developments of solid-state batteries: A review. *Materials*, 17(1), 239.

[23] Liu S, Zhou L & Neyts K, 2026. From promise to production: Strategy for halide-based all-solid-state battery pilot lines. *Advanced Energy Materials*, 16(4), e05286.

[24] Zheng Z, Xie Y, Zhang H, Yang B, Zhou J & Zhu Y, 2026. All-solid-state batteries for the grid: A realistic appraisal of challenges and opportunities. *Energy*, 345, 140229.

[25] Zhu QC, Wang ZY & Qiu L, 2026. Advancements in recycling and regeneration technologies for solid-state batteries: Challenges, strategies, and directions. *Batteries & Supercaps*, 9(1), e202500508.

[26] Karkar Z, Houache MS, Yim CH & Abu-Lebdeh Y, 2024. An industrial perspective and intellectual property landscape on solid-state battery technology with a focus on solid-state electrolyte chemistries. *Batteries*, 10(1), 24.

[27] Borthakur PP, Sarmah P, Saikia M, Hussain TA & Medhi N, 2026. Metal oxide nanomaterials for energy density improvement in lithium-ion and solid-state batteries. *Materials Proceedings*, 25(1), 17.

[28] Wu D , & Wu F, 2023. Toward better batteries: Solid-state battery roadmap 2035+. *Etransportation*, 16, 100224.

[29] Plumeyer JF, Wolf S, Dorn B, Heimes H & Kampker A, 2026. Optimization of production processes for solid-state batteries: A methodology for scaling from laboratory to mass production. *Procedia CIRP*, 138, 863–868.

[30] Kansal N & Tripathi G, 2025. Exploring the potential of flexible thin film solid-state batteries for electric vehicle. *Future Batteries*, 6, 100057.

[31] Sotskov V, 2018. Band gap prediction for inorganic crystals with machine learning. *Master's Thesis*, Lappeenranta University of Technology.

[32] Ward L, Agrawal A, Choudhary A & Wolverton C, 2016. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Computational Materials*, 2, 16028.

[33] Saccoccio M & et al., 2024. Automatic prediction of band gaps of inorganic materials using a gradient boosted and statistical feature selection workflow. *Journal of Chemical Information and Modeling*, 64(1), 125-138.

[34] Liu Z, Fan H, Ji F, Tao Y, Dong Y, Fu S & et al., 2024. Prediction of bandgap in lithium-ion battery materials based on explainable boosting machine learning techniques. *Materials*, 17(24), 6217.

[35] Rajagopal D, 2026. AI-Driven design of next-generation battery materials. *Journal of Materials Informatics*, 6(1), 45-58.